DOCUMENT RESUME

ED 436 093 FL 026 077

AUTHOR Kitao, S. Kathleen; Kitao, Kenji
TITLE Essentials of English Language Testing.
ISBN ISBN-4-268-00327-4
PUB DATE 1999-10-01
NOTE 162p.
AVAILABLE FROM Eichosha Co., Ltd., Kusaka Building, 2-28 Kanda Jimbocho, Chiyoda-ku, Tokyo 101-0051, Japan. Tel: 03-3263-1641; Fax: 03-3263-6174; e-mail: eichosha@po.sphere.ne.jp
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Communication Skills; Computer Uses in Education; Data Analysis; *English (Second Language); Evaluation Methods; Grammar; Internet; Listening Skills; Reading Ability; Reading Skills; Scoring; Second Language Learning; Speech Skills; Standardized Tests; *Student Evaluation; Test Construction; Test Content; Test Reliability; Test Results; Test Validity; *Testing; Vocabulary Skills; Writing Evaluation; Writing Skills

ABSTRACT

This book covers essential problems and basic concepts of English language testing. There are 20 chapters in four sections. Section 1, "Background," includes: (1) "Why Teachers Test"; (2) "Some Definitions"; (3) "Validity and Reliability"; and (4) "What You Should Learn about Standardized Tests." Section 2, "Preparing To Write Tests," includes: (5) "Test Specifications"; (6) "Writing a Good Test"; and (7) "Some Common Problems with Test Items." Section 3, "Writing Tests," includes: (8) "Testing Reading"; (9) "Testing Listening"; (10) "Testing Writing"; (11) "Testing Speaking"; (12) "Testing Grammar"; (13) "Testing Vocabulary"; and (14) "Testing Communicative Competence." Section 4, "The Results," includes: (15) "Scoring Tests"; (16) "Graphs and Descriptive Statistics"; (17) "Item Analysis"; (18) "Computer Programs for Analyzing Test Results"; (19) "Evaluating Test Results"; and (20) "Reporting the Results of Examinations." Each chapter also includes a list of key words, an introduction, a conclusion, and either an application exercise or discussion questions. The two appendixes present types of language testing resources on the Internet and a list of language testing resources on the Internet. (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 436 093

FL

Essentials of English Language Testing

S. Kathleen Kitao, PhD
Kenji Kitao, PhD

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
S. Kathleen Kitao
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

EICHOSHA

2

Essentials of English Language Testing

S. Kathleen Kitao, PhD
Kenji Kitao, PhD

EICHOSHA CO., LTD.
TOKYO

Copyright © 1999 by Dr. S. Kathleen Kitao and Dr. Kenji Kitao. All rights reserved. No part of this publication may be reproduced in any form without written permission from the authors and the publisher.

Cover design: Toshiyuki Takei

Printed in Japan.

Eichosha, Kusaka Bldg., 2-28 Kanda Jimbocho, Chiyoda-ku, Tokyo 101-0051
Phone 03-3263-1641 Fax 03-3263-6174
e-mail: eichosha@po.sphere.ne.jp

Preface

English teachers often make, administer and grade tests for the purpose of evaluating their students, for giving them a final grade. In many cases, tests are made for only that purpose. Teachers may spend a great deal of time and energy on testing, but they often do not use tests as effectively as they could to get feedback that would help them improve instruction and students' learning.

There are many standardized English language tests, and many people take them. They receive certain scores and interpretations for those scores, and they often tend to rely on the scores without actually understanding exactly what those scores mean.

English tests are often used in entrance examinations to screen students. There are many arguments about whether entrance examinations are good tests or not. Are they good proficiency tests? Good achievement tests? Are they good for distinguishing between students who will do well in college and those who will not? Often people who are arguing about entrance examinations are actually arguing about different questions. Those are different arguments, and when they are discussed, it is necessary to be clear about which problem is being addressed.

Testing is a complex subject. Most English teachers have not been well trained in the making, administering, grading, and interpreting of tests. They may have misconceptions about what testing is, what is being tested in a particular test, how the results of a test can be interpreted and used, and so on. In many cases, teachers make wrong decisions if they lack a knowledge of testing.

In this book, we will cover essential problems and basic concepts of testing, such as validity and reliability, different types of tests, test specifications, testing the four skills, and using test results. It is not designed for experts of testing, such as people who are making standardized tests or any large-scale tests which are given to thousands of students. It is designed for classroom English teachers and students who are going to be English teachers. It covers general information about testing that is

useful for English teachers who are teaching in any situation or at any level.

We hope that this book will help teachers to make, administer, grade, and interpret tests. We hope that it will help them to improve their instruction, so that their students will learn English better, and they can get better results from their English instruction.

If you are interested in doing further reading about testing, a bibliography is available at http://ilc2.doshisha.ac.jp/users/kkitao/library/biblio/test-bib.htm

* * * * * * * * *

We like to express our appreciation to Dr. Caroline Clapham and Dr. J. Charles Alderson, who gave us a great deal of information about testing. Without their assistance and information, this book would not have been possible. We are also grateful to the faculty of the Linguistics Department at Lancaster University, which gave us the opportunity to do research for a year. Without their assistance and facilities, we could not have written this book. Finally, we would like to thank Todd Schuett of Doshisha Women's College, Richard Hogeboom of Doshisha Women's College, and William Reis of Doshisha Women's College for reading this book and making valuable comments.

S. Kathleen Kitao, PhD
Kenji Kitao, PhD

January, 1999
Kyoto, Japan

Table of Contents

Preface ..... *iii*

Section I. Background

Chapter 1. Why Teachers Test *1*

I. Introduction *1*

II. Testing for Feedback *1*

A. Feedback about Students

B. Feedback about Instruction

The teacher — The materials

III. Testing for Motivation *3*

A. Harmful Backwash B. Beneficial Backwash

IV. Conclusion *3*

Discussion Questions *4*

Chapter 2. Some Definitions *5*

I. Introduction *5*

II. Norm-Referenced vs. Criterion-Referenced Tests *6*

A. Criterion-Referenced Tests B. Norm-Referenced Tests

III. Divisions According to Purpose *6*

A. Achievement Tests

Achievement Tests — Progress Tests

B. Placement Tests C. Proficiency Tests

D. Aptitude Tests E. Diagnostic Tests

IV. Methods of Testing *8*

A. Direct vs. Indirect Tests B. Oral vs. Written Tests

V. Other Distinctions *9*

A. Objective vs. Subjective Tests

B. Discrete Point vs. Integrative Tests

C. Speed vs. Power Tests

VI. Conclusion *10*

Discussion Questions *10*

Chapter 3. Validity and Reliability *11*
I. Introduction *11*
II. Validity *11*
III. Types of Validity *12*
A. Internal Validity
Face validity — Content validity
B. External validity
Concurrent validity — Predictive validity
IV. Reliability *14*
A. Test-Retest Reliability
Test-retest — Parallel groups — Parallel tests
B. Inter-item Consistency
Statistical tests — Split-half
C. The Relationship Between Validity and Reliability
V. Conclusion *15*
Discussion Questions *16*

Chapter 4. What You Should Learn about Standardized Tests *17*
I. Introduction *17*
II. Necessary Information about Tests *18*
A. Purpose B. Audience C. Test Specifications
D. Types of Tests E. Questions F. Administration
G. Scores H. Statistical Data I. Reliability
J. Previous Records K. Other Information
III. What Teachers Should Do *20*
A. Contact Information
IV. Conclusion *21*
Application Exercise *21*

Section II. Preparing to Write Tests

Chapter 5. Test Specifications *22*
I. Introduction *22*
II. Content of Test Specifications *23*
III. Using Specifications for Classroom Tests *25*

IV. Conclusion *26*
Application Exercise *26*

Chapter 6. Writing a Good Test *28*
I. Introduction *28*
II. Goals *28*
III. Audience *29*
IV. Weighting *29*
V. Clarity *30*
VI. Length *30*
VII. Simplicity *31*
VIII. Objectivity *31*
IX. Practicality *32*
X. Economy *32*
XI. Scorability *32*
XII. Pretesting *33*
XIII. Conclusion *33*
Discussion Questions *33*

Chapter 7. Some Common Problems with Test Items *34*
I. Introduction *34*
II. Item types *34*
A. Grammar Items B. Vocabulary C. Pronunciation
D. Functions E. Reading
III. Format Types *38*
A. Multiple Choice Questions B. True/False Questions
IV. Conclusion *40*
Application Exercise *41*

Section III. Writing Tests

Chapter 8. Testing Reading *42*
I. Introduction *42*
II. Skills Involved in Reading *42*
III. Choosing a Text *43*

A. Variety B. Type of Text
C. Reading and Background Knowledge D. Difficulty
IV. Tasks to Test Reading *45*
A. Testing Low Level Skills
Word recognition — Sentence recognition — Word and picture matching
B. Tasks for Middle and Higher Level Students
True/false questions — Multiple choice questions —
Short answer/completion questions — Ordering tasks —
Charts — Main idea
V. Conclusion *49*
Application Exercise *49*

Chapter 9. Testing Listening *50*
I. Introduction *50*
II. Testing Phoneme Discrimination *50*
III. Discriminating Stress and Intonation *51*
IV. Understanding Sentences and Dialogues *52*
V. Tasks Using Visual Materials *53*
A. Matching and True/False Tasks B. Map Tasks
C. Drawing Tasks
VI. Tasks Involving Talks and Lectures *54*
VII. Conclusion *55*
Application Exercise *56*

Chapter 10. Testing Writing *57*
I. Introduction *57*
II. Components of Writing *58*
III. Types of Writing Tasks *58*
A. Gap Filling B. Form Completion C. Making Corrections
D. Letter Writing E. Essay Writing
IV. Issues Related to Intermediate and Advanced Writing Tests *60*
A. Instructions
Time and length — Establishing a context and communicative
purpose

B. Choice of Topic(s)
Doability — Knowledge — Motivation — Breadth — Allowing students to choose topics
C. Choice of Task(s)
Connected discourse — Realistic task — Clarity — Modes of discourse — The number of tasks
D. Miscellaneous
Level of difficulty — Time allowed
V. Marking Tests of Writing *63*
VI. Conclusion *63*
Application Exercise *64*

Chapter 11. Testing Speaking *65*
I. Introduction *65*
II. Difficulties of Testing Speaking *65*
III. Types of Speaking Tests *66*
A. Reading Aloud B. Conversational Exchanges
C. Testing Speaking Using Visual Material
D. Oral Interviews E. Role Play Test
F. Group or Pair Activities
IV. Assessment *70*
V. Conclusion *71*
Application Exercise *72*

Chapter 12. Testing Grammar *73*
I. Introduction *73*
II. Types of Tests *73*
A. Multiple Choice Tests B. Error Recognition
C. Items to Test Knowledge of Word/Sentence Order
D. Completion Items E. Transformation Items
F. Word Changing Items G. Sentence Combining Exercises
III. Conclusion *78*
Application Exercise *78*

Chapter 13. Testing Vocabulary *79*
I. Introduction *79*
II. Types of Vocabulary Knowledge *79*
III. Sources of Words to Test *80*
IV. Multiple Choice Tasks
A. Synonym/Definition Items B. Picture Items
C. Sentence Items D. Context Items
V. Set Items *83*
VI. Matching Items *84*
VII. Completion Items *84*
VIII. Word Formation Items *84*
IX. Guessing Meaning from Context *85*
X. Conclusion *86*
Application Exercise *86*

Chapter 14. Testing Communicative Competence *87*
I. Introduction *87*
II. The Communicative Approach to Testing *88*
A. What Communicative Language Tests Measure
B. Tasks C. Evaluation
III. Examples of Communicative Test Tasks *90*
A. Speaking/Listening
Information gap — Role play
B. Testing Reading and Writing
Letter writing — Summarizing
C. Testing Listening and Writing/Note Taking
IV. Conclusion *93*
Application Exercise *94*

Section IV. The Results

Chapter 15. Scoring Tests *95*
I. Introduction *95*
II. Machine Grading *96*

III. Hand Grading *97*
A. Objective Tests B. Subjective Tests
IV. Adjusting Scores *98*
V. Conclusion *98*
Application Exercise *98*

Chapter 16. Graphs and Descriptive Statistics *99*
I. Introduction *99*
II. Graphs *99*
A. Histograms B. Polygons C. Pie Charts
III. Descriptive Statistics *100*
A. Measures of Central Tendency
The mean — The mode — The median
B. Measures of Dispersion
Range — Standard deviation
IV. Conclusion *102*
Application Exercise *102*

Chapter 17. Item Analysis *105*
I. Introduction *105*
II. Facility Value *105*
III. Discrimination Index *106*
IV. Looking at Incorrect Alternatives *107*
V. Examples *108*
VI. Other Types of Items *109*
VII. General Comments *110*
VIII. Conclusion *110*
Application Exercise *110*

Chapter 18. Computer Programs for Analyzing Test Results *115*
I. Introduction *115*
II. Types of Programs Available *116*
A. Spreadsheets B. Iteman C. SPSS and SAS D. Databases
E. Test Generating Programs F. Word Processors G. E-mail
H. Lists

III. Conclusion *118*
Application Exercise *119*

Chapter 19. Evaluating Test Results *120*
I. Introduction *120*
II. How to Evaluate Test Results *120*
III. Giving Feedback to Students *121*
IV. Evaluating Instruction *122*
V. Improving Tests *122*
VI. Conclusion *123*
Application Exercise *123*

Chapter 20. Reporting the Results of Examinations *124*
I. Introduction *124*
II. General Considerations *124*
III. What Do You Report to Different Audiences? *126*
A. Teachers B. Students and their Parents
C. Administrators D. General Audience
E. Specialists in Testing
IV. Conclusion *129*
Discussion Questions *129*

Appendix A
Types of Language Testing Resources on the Internet *130*
I. Introduction *130*
II. Language Testing Resources *130*
III. Conclusion *132*
Application Exercise *132*

Appendix B
A List of Language Testing Resources on the Internet *133*

Section I
Background

Chapter 1
Why Teachers Test

Key Words

feedback
motivation
harmful backwash
beneficial backwash

Introduction

Evaluation is required by most language programs, so it is often necessary to give tests. Though teachers may have no choice but to give tests, there are also good reasons to do so, and if the tests and their results are used wisely, there are even benefits of testing, both for the students as individuals and for the program as a whole. In this chapter, we will discuss reasons for testing and how testing can be used to benefit both the student and the teacher.

Testing for Feedback

Feedback about Students

Teachers use tests, most obviously, to get feedback about students' abilities for the purpose of evaluating the students. There are different types of tests for different purposes. One purpose is to see how much students have achieved, usually in order to assign a grade. This indicates

how much the students have learned in the class. Another purpose is to place students in the appropriate class or group, based on their performance on the test. Language tests are used to see how proficient students are in the language they are studying, for example, to see whether a company employee can function in an English-speaking country. Tests are also used to determine the aptitude of a particular student for learning languages in general. Finally, tests are used to diagnose the specific problems that a student is having in learning a language.

However, if tests are used only to evaluate the students in this way, they are being underused. Tests can also be used to give students feedback. A frustration of studying a language is that it is sometimes difficult to see progress. Tests can be used to allow students to see how well they are doing and how much progress they have made. They can also let students know what areas they are weak in so that they can put more work into those areas. Also, using this feedback, the teacher can see what problems individual students have and help them with those problems.

Feedback about Instruction

The teacher. In addition to using tests to get feedback about students, a teacher can use tests to get feedback about his or her own teaching. Tests allow teachers to see how students are doing in order to allow them to improve their own effectiveness and to make adjustments in their teaching. Tests help the teacher see exactly what the students' areas of difficulty are and guide them in giving the students remedial work or additional practice on those areas.

The materials. Testing may also help the teacher see weaknesses in the materials, that is, it helps the teacher recognise where explanations are unclear or otherwise insufficient, where not enough practice of a particular point is given. If the teacher finds some insufficiency in the materials, he/she can supplement them with additional explanations or exercises. If most or all students do poorly on a test, the teacher should consider whether the material being covered in the class is too difficult and adjust it accordingly. Tests can give the teacher an idea of whether new material is being introduced too quickly or too slowly.

Testing for Motivation

Tests can also motivate students, an effect that is called backwash. Backwash refers to the fact that the test influences the learning of the students. When students know that they will be tested, that knowledge influences what they study and how they study, because studying certain material or in a certain way will help them do well on the test. Backwash has the potential to be either beneficial or harmful.

Harmful Backwash

If the test becomes the focus of the class, and if the test is divorced from the real use of the language, this can cause harmful backwash. For example, if writing ability is tested using a multiple-choice test rather than actual writing, the students will concentrate on skills that have nothing to do with writing. Similarly, if pronunciation is tested with a written test using the phonetic alphabet, the students will concentrate on learning the phonetic alphabet, and they may be able to do well on the test, even though they cannot actually pronounce English well. Such tests can have a harmful effect on learning.

Beneficial Backwash

In contrast, if a writing test involves applying actual writing skills, that will cause beneficial backwash. The skills that students will be developing in order to take the test will also be skills that are useful to them in writing. In other words, if tests are designed wisely, they can motivate students to do the kinds of things that they ought to be doing in order to learn to communicate effectively in a language. They can help the students set concrete goals for their study. If tests are poorly designed, they motivate students to concentrate on the test itself and not on communicating in the language.

Conclusion

Tests are useful for much more than giving grades to students. They are also important for giving students feedback about how they are doing, and for giving the teacher feedback about how individual students

and the class as a whole are doing. This feedback is valuable in improving instruction and supplementing materials as necessary.

Discussion Questions

1. Think about language tests that you have taken in the past. Briefly describe a few of the tests.
2. How did you feel about taking the tests? Did you experience any benefits from them? If so, what?
3. Did you feel they were harmful to your language learning in any way? If so, how?
4. Did you feel that you were being evaluated fairly as a result of the tests?

Chapter 2

Some Definitions

Key Words

norm-referenced tests
criterion-referenced tests
achievement tests
progress tests
proficiency tests
aptitude tests
diagnostic tests
direct tests
indirect tests
oral tests
written tests
objective tests
subjective tests
discrete point tests
integrative tests
speed tests
power tests

Introduction

Language tests come in many different types, and they have different purposes and can be used in different ways. In order to understand how to construct a test and how to use the results, it is necessary to understand the different types of tests and their characteristics. In this chapter, we will discuss these.

Norm-Referenced vs. Criterion-Referenced Tests

Various characteristics are used to distinguish among types of tests, but one important distinction is made according to the standards used to grade the tests. This distinction is between norm-referenced tests and criterion-referenced tests. Essentially, norm-referenced tests compare students to other students; criterion-referenced tests compare the students to some standard.

Criterion-Referenced Tests

Criterion-referenced tests are tests based on particular, defined descriptions of the language proficiency expected. All test-takers are judged against that standard, rather than against other test-takers. These tests are generally used for testing speaking and writing. Scales are provided for the examiners, which describe the standard of language expected at each level, and the test-taker is assigned to the level that seems to best match his/her performance. All test-takers who achieve a certain level are evaluated as having that level, regardless of how many other test-takers achieved that level.

Norm-Referenced Tests

In the case of norm-referenced tests, students are compared either to other students in the same class or to other students who have taken the test at different times and in different places. Reporting results in terms of percentiles indicates a norm-referenced test, since percentiles indicate how many test-takers did better than a particular test-taker and how many did worse. The Test of English as a Foreign Language (TOEFL) is a norm-referenced test.

Divisions According to Purpose

Tests can also be classified according to their purpose. Among the common types are achievement tests, placement tests, proficiency, and aptitude tests. These tests have different characteristics, and the material for these tests also come from different sources.

Achievement Tests

Achievement tests. An achievement test is a test based on past learning, and so it tests how much the student has learned. Achievement tests are often given at the end of a course to see how much of the material covered in the course the student has mastered. The content is drawn from the syllabus of the course the students are taking, from the textbook, from homework assignments, etc.

Progress tests. Progress tests are a specific type of achievement test. They are achievement tests that are given periodically during a course, measuring whether a student is learning what is being covered in the course. The content is also drawn from what is covered in the course.

Placement Tests

Placement tests are tests that are used to decide at which level of a course a student should be placed. These tests may, like achievement tests, draw on the content of the courses, because they are intended to determine what part of the content of the courses students have already covered and what part they still need to learn, or they may test more general language. In some cases, students may be placed differently according to their performance on tests of different skills. That is, a student who does well on the reading section of the test would be placed at a high level in reading, but if the same student does poorly on the listening section, he/she will be placed at another level in listening.

Proficiency Tests

Proficiency tests are intended to test the language ability of test-takers of different language training backgrounds in order to see whether the test-takers have reached a given level of language ability. Proficiency tests can be divided into general and special purpose tests. General tests cover the language in general rather than being drawn from a particular pool of language. Special purpose tests are tests that deal with the language needed for a specific purpose, for example the language that a doctor or a diplomat would need. These are based on specifications of language needed, usually based on needs analysis for the particular purpose of the type of people who would be taking the test.

Aptitude Tests

Another type of test is an aptitude test. Aptitude tests measure the ability to learn a foreign language. They may test such skills as intelligence, memory, deduction of rules, and other skills that seem to be related to the learning of language. Aptitude tests are related to the skill of language learning in general, not to a specific language, for example, by testing the student's ability to deduce grammatical rules from examples.

Diagnostic Tests

Diagnostic tests are used to determine what particular problems students have. This is actually difficult to do, since it is difficult to divide language ability into component skills.

Methods of Testing

Tests can also be distinguished according to the method used for testing, both whether the skill is tested directly or indirectly and whether it is tested orally or on paper.

Direct vs. Indirect Tests

Tests can be direct and indirect. In a direct test, the test-taker actually performs the action being tested. A direct test is as much like a real-life situation as possible. For example, a direct test of English might be a role play. An indirect test is a test that tests knowledge about the action, but it does not test the action itself. For example, a multiple choice grammar test would test knowledge about grammar, but it does not test whether the test-taker can use the grammar in actual communication. It is preferable that tests be as direct as possible, but this is not always feasible.

Oral vs. Written Tests

Language can be tested orally or in writing. There are a great many types of written tests, including true/false, multiple choice, short answer, fill-in-the-blank, and essay tests. Oral tests can also be done in a variety of ways, including interviews, role plays, listening and answering questions, etc. All of these methods of testing have advantages and disadvantages. They are also appropriate for testing different skills and in differ-

ent situations. The teacher should be aware of these in choosing a method of testing.

Other Distinctions

In addition to the above types of language tests, there are various other distinctions among tests.

Objective vs. Subjective Tests

Objective tests are tests in which there are "right" and "wrong" answers. The task of the testee is to find the "right" answer, and the marker of the test has little discretion in making a decision about the answer. Multiple-choice comprehension questions about a reading passage would be objective questions. In a subjective test, the answers are not right or wrong. The testee tries to apply principles that have been learned, and the marker has to decide to what extent the answer matches particular standards. An essay in response to a question would be a subjective test.

Discrete Point vs. Integrative Tests

Another distinction made in testing is between discrete point and integrative tests. Discrete point tests are tests of one particular skill or piece of knowledge. Integrative tests, in contrast, test two or more skills at once.

Speed vs. Power Tests

Finally, speed tests are tests in which students are given only a limited amount of time to finish the test, so they have to complete the test as quickly as possible. This reflects real situations, because the test-takers may have a limited time to perform language tasks in real life situations. In this type of test, the items are generally relatively easy, so that given enough time, most students should be able to answer all of the questions. Power tests are tests in which test-takers have as much time as they need to finish the test. However, the test items are more difficult than in a speed test, so not all of the students can answer all of the questions.

The significance of speed and power tests for language testing in particular is that speed tests can be used to test fluency, while power tests can be used to test accuracy. Fluency might be said to correspond with

performance and accuracy with competence. Judicious use of power and speed tests can help the teacher and student see what the problem areas are in terms of fluency and accuracy so that they can work on them.

Conclusion

There are many different types of language tests. In this chapter we have discussed tests for different purposes as well as the differences between norm-referenced and criterion-referenced tests, direct and indirect tests, discrete point and integrative tests, and speed and power tests. In making or choosing tests, it is crucial to be aware of these different types of tests and the options that they allow you in order to meet the goals that you have set for the tests.

Discussion Questions

A. Bring a test that you have taken recently (or look at a test that your teacher has provided) and answer the following questions about it.

1. Was it a norm-referenced or criterion-referenced test? Why? What were the relevant norms or criteria?
2. Was it an achievement, placement, proficiency, diagnostic or aptitude test? Why?
3. Was it a direct or indirect test? Why?
4. Was it a subjective or objective test? Why?
5. Was it a discrete point or integrative test? Why?
6. Was it a power or speed test? Why?

B. Decide on a situation in which you might give a test (or consider the situation that the teacher assigns to you). Who are the students? What is their level of proficiency? What is the purpose of the test? Considering the various definitions discussed in this chapter, which kind of test would you use? (Direct or indirect, objective or subjective, discrete point or integrative, and so on?) Explain your reasons.

Chapter 3
Validity and Reliability

Key Words

validity
internal validity
face validity
content validity
external validity
concurrent validity
predictive validity
reliability
test-retest reliability
parallel groups
parallel tests
inter-item consistency

Introduction

When making a test, there are two basic factors to consider—validity and reliability. In this chapter, we will discuss what these are and how they must be considered.

Validity

Validity can be defined as the degree to which a test actually tests what it is intended to test. If the purpose of a test is to test ability to communicate in English, then it is valid if it does actually test ability to communicate. If what it is testing is actually knowledge of grammar, then it is not a valid test *for testing ability to communicate.*

This definition has two very important aspects. The first is that validity is a matter of degree. A test does not have to be evaluated as completely valid or completely invalid; some tests are more valid than

others. A second important aspect of this definition is that tests are only valid or invalid in terms of their intended use. If a test is intended to test reading ability, but it also tests writing, then it may not be as valid for *testing reading as for testing reading and writing together.*

Types of Validity

Validity is divided into different types. Broadly, it is divided into internal and external validity.

Internal Validity

Internal validity is validity in terms of the test itself—whether the content of the test and the way the test is carried out allows it to test what it is intended to test. There are two types of internal validity—face validity and content validity.

Face validity. Face validity is the extent to which a test actually appears to test what it is intended to test. Thus it is the impression a test makes on general readers as to what skills it requires. While such opinions are not expert, they can be useful, because they indicate the kind of response that you can get from the people who will take the test. If a test does not appear to be valid to the test takers, they may not do their best, and in this way, the perceptions of non-experts can be helpful.

Content validity. Content validity is the opinion of experts as to whether the material covered in a test is a valid sample. The experts should look at whether the test is representative of the skills you are trying to test. This involves looking at the syllabus (in the case of an achievement test) and the test specifications and deciding what the test intends to test and whether it accomplishes what it intends to. The problem is that even experts may disagree over the content validity of a test, particularly if they are not given a systematic way of looking at it. However, even looking at a test systematically does not guarantee that all the experts will agree on its relative content validity.

External Validity

External validity has to do with the relationship between the test and other measures. There are two types of external validity—concurrent

validity and predictive validity.

Concurrent validity. Concurrent validity is the degree to which a test correlates with other tests that test the same thing. In other words, if a test is valid for a certain purpose, its results should be similar to the results of other measures that are valid for the same purpose. When considering concurrent validity, there are two concerns.

First, the measure that is being used for comparison of the test in question must be valid. If that measure is not valid, there is no point in using it as a standard of comparative validity. For example, a teacher's ranking might be used to establish validity—but the teacher's ranking may be influenced by a number of factors that are not related to the students' actual proficiency. However, if you must use such a measure, one possible approach is to average the rankings of several teachers to try to compensate for any individual bias.

Second, the measure must be valid for the same purpose as the test whose validity is being considered. A reading test cannot be used to gauge the concurrent validity of a grammar test. In addition, if teachers' rankings are being used, it is necessary to make certain that they understand on what basis they are expected to rank the students. If the test being considered is a grammar test, then the teachers should be asked to rank the students according to their grammar proficiency, not their overall English language ability.

Predictive validity. Predictive validity is the extent to which the test in question can be used to make predictions about future performance. For example, does a test of English ability accurately predict how well students will get along in a university in an English-speaking country? There are a number of problems with trying to answer such a question. A measure such as students' subsequent grade point averages is sometimes used to study the predictive validity of such a test, but there are many factors other than English proficiency involved in academic success. Therefore, there is a limit to how much of subsequent performance can be predicted by even the most valid test. In addition, it is not possible to know how the students who scored too low on the tests to be admitted to the university would have done if they had been admitted.

Reliability

There are two types of reliability. Test-retest reliability is the extent to which the test would produce the same result time after time. For example, if a ruler is used to measure a piece of paper, it should get the same result every time. Therefore a ruler is a reliable measure.

One of the basics of reliability is that, all else being equal, the more items there are on a test, the more reliable it will be. With few items, one bad item can have a greater influence on the results of a test than when there are many items.

Another type of reliability is inter-item consistency. Inter-item consistency is the extent to which all the items on the test are measuring the same thing.

Test-Retest Reliability

Determining test-retest reliability is not a simple matter. There are various ways of trying to measure it, but each of them has potential problems.

Test-retest. One way of measuring reliability is to give the students the same test twice to the same group of students. However, if a test is given twice, particularly if there is not much time between the two tests, the students might do better the second time due to a practice effect. On the other hand, if there is a longer time between the two tests, the practice effect is not as likely to be important, but with the passage of time and more opportunity to learn, the students' English proficiency might have improved.

Parallel groups. Another way to measure reliability is to have two parallel groups, that is, two groups matched by ability levels, take the same test and to determine if students with matching ability levels get the same scores. The problem is determining whether the two groups are truly parallel.

Parallel tests. Reliability can also be measured by giving the same group of students parallel tests, that is, two similar tests with the same type and number of items, the same instructions, etc., and checking how well the results agree. The problem with this approach is determining

whether the two tests are actually parallel.

Inter-Item Consistency

Inter-item consistency is usually determined using statistical tests.

Statistical tests. The Kuder Richardson formulas are statistical tests that measure reliability in terms of whether all of the items in a test are measuring the same thing. Kuder Richardson 20 looks at the result of all possible divisions of the test in order to determine its inter-item consistency. (The Kuder Richardson alpha is the same, except that it is specifically used for multiple choice tests.) Kuder Richardson 21 is similar to Kuder Richardson 20, except that it takes into account the mean and standard deviation but not all individual items, which produces a lower reliability. Reliability is reported with numbers between 0.00 and 1.00. The higher the number, the more reliable the test.

Split-half. Another way of measuring inter-item consistency is to randomly assign test items to two groups and compare students' scores on these two groups of items. Of course, it is still possible that the two resulting tests will not be parallel. In addition, because the individual tests are shorter, they will be less reliable, which needs to be compensated for statistically, that is by calculating how reliable the test would have been with twice the number of items.

The Relationship Between Validity and Reliability

Validity and reliability have a complicated relationship. For a test to be valid, it must also be reliable. A test that gives different results at different times cannot be valid. However, it is possible for a test to be reliable without being valid. That is, a test can give the same result time after time but not be measuring what it was intended to measure.

Conclusion

Validity and reliability are very important to a test. Validity is the extent to which a test tests what it is intended to test; reliability has to do with whether the test produces consistent results.

Discussion Questions

Bring a test (or look at one provided by the teacher) and answer the following questions.

1. What appears to be the goal or goals of this test?

2. How does the test go about meeting the goal or goals?

3. Do you think the test is successful in meeting the goals?

4. Do you think the test has face validity?

Chapter 4

What You Should Learn about Standardized Tests

Key Words

purpose
audience
test specifications
types of tests
questions
administration
scores
statistical data
reliability
previous records
other information

Introduction

There are a great many standardized tests available throughout the world, including the Test of English as a Foreign Language (TOEFL), the Test of English for International Communication (TOEIC), the International English Language Testing System (IELTS), and the Cambridge University Certificate Examination of English. You may have access to certain standardized scores for your students, and you may wonder what the scores really mean. Your students may bring you the scores of standardized tests and ask you questions about what their scores mean or ask advice about how they can improve their scores. Students may also ask which test they should take for a certain purpose. In these cases, you will need to know what the purposes of these tests are and what the results mean.

English language teachers should be aware of common standardized

tests: what they are like, what they are intended to test, and generally what their scores mean. However, it is not possible to know all standardized tests, since there are so many of them. In addition, those tests may be changed in the near future, so even if you are familiar with the current ones, the situation may change. However, you should be aware of what the major tests are and know where to get information about tests and how to interpret the information you get. In this chapter, we will discuss those issues.

Necessary Information about Tests

In order to understand a standardized test, you need to obtain the following information. Some testing organizations have literature that gives you this information, but some do not. If the company does not have such literature, you should write to them and ask them for the information.

Purpose

Does the test have a specific purpose or is it a general proficiency test? Is it testing general proficiency, or is it testing a specific skill or competence? How many levels of proficiency do the test results measure?

Audience

Who are the intended test-takers? Is it designed for a particular type of testee or for anyone?

Test Specifications

What are the test specifications? What skill(s) do they indicate will be tested? How many questions are there on the test, how many areas will be covered, how many sections are there, what are the formats, etc.

Types of Tests

Is the test norm-referenced or criterion-referenced? Is it objective or subjective? Is it a multiple-choice test? Does it require any production, either speaking or writing?

Questions

Are any sample questions from the test available? Have previous versions of the test been published? If so, who is the publisher? What kinds of questions are included? What are they testing? What is covered by those questions? Have they changed formats or types of questions much in the past? Do they have face validity?

Administration

Who is responsible for administering the test? How often is it administered? Where is it administered? When is the deadline for applications? Where do people apply for the test? How much is the fee? How long is the test? Are there time limits for individual sections? If so, what are they? How soon do the test administrators report the results to the testees?

Scores

Do the administrators give scores to individual testees? Do they report scores for each section or only the total score? Do they explain how the final score was calculated for that particular test? Do they give explanations for the score? Do they just inform testees whether they passed or not? If so, do they inform testees what the cut-off point for a passing score is and how far the testee was from it? Do they inform testees how the cut-off point was decided? Do they give any suggestions to improve the score in the future?

Statistical Data

Do the administrators give any statistical data about the test results, such as mean, median, mode, standard deviation, standard error, z scores, score distributions, etc.? (See Chapter 16 for further explanation.) Do they release any evaluation of or comment on the results?

Reliability

Do the administrators publish the reliability of the test? If they do, what is the reliability? Do they publish correlations with other well known standardized tests? If so, with which tests do they show correlations? What are the correlations?

Previous Records

Do the administrators of the test publish records of previous tests? If so, what kinds of records? Do they publish the number of test takers? Do they publish their score distributions? Do they indicate anything about the problems of the previous testees? What else do they publish?

Other Information

Is there any other information about the test? What information is available? What do you understand from such information?

What Teachers Should Do

English teachers should collect the above information whenever it is available and file it. The information should be classified according to whether it came from the test administrators or from a third party. In the process of gathering such information, you can categorize standardized tests into several categories by purposes, levels of English, etc.

It is always good to talk with your students about their experiences taking certain tests. Their opinions are, of course, not the opinions of experts, but as the people that the tests are intended for, their feedback can be revealing.

With such knowledge, you can discuss with your students which tests they should take, what their scores on the tests indicate, and what kinds of preparation they need, etc. You will then be more capable of helping your students, and if many of your students are taking the test, you will know how to improve your instruction for those students.

Contact Information

The following is contact information for some major tests.

<u>IELTS</u>

University of Cambridge
Local Examinations Syndicate
1 Hills Road
Cambridge, CB1 2EU
United Kingdom

Tel.: (UK) 01223 553311 (International) +44 1223 553311
Fax: (UK) 01223 460278 (International) +44 1223 460278
E-mail: helpdesk@ucles.org.uk
Web site: http://www.ielts.org/

TOEFL

Corporate Headquarters
Educational Testing Service
Rosedale Road
Princeton, NJ 08541 USA
Tel.: 609-921-9000 Fax: 609-734-5410
E-mail: etsinfo@ets.org
Web site: http://www.toefl.org/

TOEIC

International Communications Incorporated (ICI)
Address: International Communications Incorporated (ICI)
3301 Country Club Road, Ste 2205
Endwell, NY 13760
Tel.: 607-748-9500 Fax: 607-748-9614
E-mail: USICI@AOL.COM
Web site: http://toeic-usa.com/

Conclusion

It is very important to make, administer, grade, interpret, and improve your own tests. It is also important to understand standardized tests and interpret their scores correctly, using such information as who the prospective testees are, what type of test it is, what type of questions the test includes, and so on. Having this information will help you help your students to take such tests more effectively.

Application Exercise

Look at the information your teacher gives you about a standardized test. Try to find the answers to the questions in this chapter.

BEST COPY AVAILABLE

Section II
Preparing to Write Tests

Chapter 5
Test Specifications

Key Words

purpose
learner
sections
text types
language skills
language elements
tasks
number of items
weight
test methods
instructions
criteria for assessment

Introduction

Test specifications are an official statement of what the test covers and how it covers it. In other words, it is a description of the test, written beforehand, which includes such information as what the purpose of the test is, what will be covered by the test, how it will be tested, how many items there will be and of what type, etc. Test specifications are written before the test, and then the test writers base the test on the specifica-

tions. After the test is finished, the specifications should be consulted, preferably by someone not directly connected with the writing of the test, to see whether they match the test that was actually written. This process is normally done for large standardized tests, but it is also a useful practice for individual teachers making tests for their own students. Test specifications can help the teacher clarify whether the test covers what it is intended to cover in the way it is supposed to cover it. In this chapter, we will discuss what is included in test specifications and how they should be used.

Content of Test Specifications

Test specifications in general cover the following areas, as appropriate. Test specifications naturally vary according to their uses, so not all of these will be appropriate for all tests.

1. What is the purpose of the test?

The purpose of a test generally falls into one of five broad categories: placement, progress, achievement, proficiency, and diagnostic. (These purposes are discussed in more detail in Chapter 2.) It is important before starting to write a test to know which of these broad purposes the test has.

2. What sort of learner will be taking the test?

Useful information about the learner can include the age or educational level; general level of proficiency; first language(s); cultural or national background; level and nature of education; reason for taking the test; professional interests, if any; and levels of background knowledge.

3. How many sections should the test have, and how long should they be?

The specifications should establish how many sections the test has, how long each of them is, and how they are different. For example, the test might be one two-hour exam or two one-hour sections, one an examination and one an essay.

4. What text types should be used in the test?

The specifications should indicate whether the texts should be

written or spoken, what kinds of sources they should come from, what topics they should include, how difficult they should be, what their functions should be (for example, persuasion or summarizing), etc.

5. What language skills should be tested?

The specifications should indicate what skills the test should cover and whether they should be tested in an integrative or discrete way. They should also establish whether the test should ask for the main idea, specific details, inferences, etc.

6. What language elements should be tested?

If there are specific grammatical points, functions, or lexical items that should be covered in the test, the specifications should list these.

7. What sort of tasks are required?

The specifications should indicate whether the tasks should be simulated authentic tasks, objective or subjective, etc.

8. How many items are there in each section, and what is the relative weight for each item?

The specifications should specify the number of items in each section and indicate whether they are weighted equally or whether more weight is given to more difficult or longer items.

9. What test methods are used?

The test specifications should indicate whether the items should be multiple choice, fill-in-the-blank, picture description, role play using cue cards, essay, etc.

10. What instructions should be given to the candidates?

The test specifications should indicate what information should be included in the instructions, whether examples of worked problems should be provided, and whether there will be information about how the responses will be evaluated.

11. What criteria will be used for assessment?

The specifications should establish whether the test will be assessed according to accuracy or fluency, whether spelling will be counted, and

so on.

Using Specifications for Classroom Tests

A classroom teacher may not write formal specifications for achievement or progress tests given in the class. However, there are good reasons to write at least informal specifications, even for classroom tests.

A classroom test should be closely related to the material covered in class. In addition, the weight given to each section should reflect what the students were intended to learn as well as the amount of time spent on it in class. Specifications can help with this, because the specifications can include what was covered in the class and what percentage of the class time was spent on it. Therefore, one part of the specifications should be a list of the areas covered in class and the weight, in terms of percentages given to them. For example, if the course was organized in terms of grammatical structures, part of the list might be:

1. present continuous tense 5%
2. articles 5%
3. countable and uncountable nouns 10%

If, on the other hand, the teacher takes a communicative approach emphasizing functions, part of the list might be:

1. greetings 5%
2. invitations 10%
3. asking for information 15%

In other words, if the course takes a structural approach, the specifications, and consequently the test, should reflect that. If, on the other hand, the course emphasizes functions, the specifications and, as a result, the test, should also emphasize functions. It is always tempting to emphasize the parts of the course that are easiest to test, rather than the parts that are important to test. Writing up test specifications listing what was covered in the class and what weight it was given can help avoid this.

Conclusion

Obviously, not everything covered in a class, or everything that students know about the language they are studying, can be covered in a test. However, the test specifications help the teacher make principled decisions about what to include in the test.

Application Exercise

Read the following test specifications. Is everything included that should be included in test specifications? Could you write a test from these specifications? If not, what other information would you need?

The test is intended to test the material covered in the first four weeks of a grammar course. It is a written test that should be thirty minutes long.

The Test

The focus of the test is on knowledge of grammar. Candidates have to demonstrate an ability to recognize and produce correct grammar. The content of the questions should be general and not tied to a particular content area.

Grammar Points To Be Tested

The test should cover the following grammar points, with approximately the same amount of emphasis on each.

1. simple past tense
2. "a," "some," and "any"
3. "how much" and "how many"
4. present continuous tense
5. past continuous tense
6. adverbs of frequency
7. "if" and "unless"
8. countable/uncountable nouns

Item Types

The test should be objectively gradable. Items can include, but not be limited to, multiple choice questions, error correction questions, gap filling, and matching. There should be at least three sections to the test, with a different type of item for each section. Each grammar point should be represented in each section.

Instructions

Instructions should be provided for individual sections. If necessary, examples should be provided.

Chapter 6
Writing a Good Test

Key Words

goals
audience
weighting
clarity
length
simplicity
objectivity
practicality
economy
scorability
pretesting

Introduction

What makes a test good or bad? The most basic and obvious answer to that question is that good tests measure what you want to measure (that is, are valid), and bad tests do not. This is discussed in more detail in the chapter on validity. However, other factors are important in making a good test. In this chapter, we will discuss some characteristics of tests that need to be considered in making, administering, and evaluating them.

Goals

In writing a test, you must always keep in mind the goal of the test. The goal of the test is what you want to measure. To the greatest extent possible, the test should measure just what you want it to measure and nothing else. (However, if you are trying to test English communicative-

ly, it is necessary to test more than one skill at once.) If you are trying to test one skill or area of knowledge, you should avoid a method of testing that requires another skill or area of knowledge. For example, if you want to test your students' knowledge of vocabulary, you need to make sure that it does not require a higher knowledge of grammar than students are likely to have. If it does, then the test is testing something other than vocabulary knowledge. You will not know whether a student misses a question due to lack of knowledge of the vocabulary item or lack of knowledge of the grammatical form.

Audience

You need to consider your audience carefully. In language testing, the level of difficulty of the language tested is very important. If you have a homogeneous group, you need a test that is the right level for that group. If it is too difficult for the group, all of them will do poorly, and you will not be able to determine anything about their knowledge or skill and how they differ in that knowledge or skill. If the group is heterogeneous, the problem is more difficult. You need to include some easy and some difficult questions, so that every testee has some questions within his/her ability level and every testee has some questions that are challenging.

Getting the level of the test right can be difficult. If possible, you should pre-test the test with students similar to the ones who will be taking it. However, this is not always practical.

Weighting

In writing a test, you need to consider how much weight each item, each section, each skill, etc., is given. It seems obvious to say that you should give the greatest weight to the most important things you are testing, but this does not always happen. It is tempting to devote the largest part of the test to whatever is easiest to test, not to what is most important to test. Therefore, the test maker needs to carefully consider what should be tested, depending on the syllabus of the class or other bases for the test, and make sure that the weighting reflects these. Test specifications

are useful here, because they allow the test maker to compare what should be covered on the test with what is actually covered on the test.

Clarity

In order to get a true idea of a student's ability, it is necessary to make sure that the student understands what is expected on the test. Obviously, if the student did not understand what to do, you cannot judge whether he/she could have done it successfully. You cannot tell whether incorrect answers were the result of not knowing the information or of not understanding the instructions. It is absolutely necessary to give clear, concise instructions. It may also be useful to provide an example of a correctly answered problem, which helps the students understand exactly what is expected. It is especially important to pre-test the instructions, or at least have someone else read them to make certain they are clear. What may seem to be perfectly clear to the writer of the instructions may be ambiguous or otherwise unclear to someone else.

Length

The more items a test has, the more reliable it is. That is, on a short test, a few wrong answers can have a great effect on the overall results, but on a long test, a few wrong answers will not influence the results as much. However, in practice, a long test has drawbacks as well. If a test is too long, and particularly if students are doing the same kind of item over and over, they may get tired and not respond accurately or seriously. If there are reasons why a test needs to be long, it is a good idea to divide the test into sections with different kinds of tasks, so that the variety of tasks will help students maintain their concentration.

Simplicity

There are many ways to measure language-related skills, and obviously some are more complicated than others. If you choose a more complicated way to test English, there can be a greater likelihood of errors or a greater likelihood that the test is testing something other than what it is

intended to test. Thus, it is important to keep the method of testing as simple as possible, while still testing the skill you intend to test. For example, a logic problem might be used to test reading ability, and the teacher might conclude that if the student got the right answer on the logic problem, he/she had good reading comprehension. However, such a test would also be testing the student's ability to work the logic problem, a skill that is not related to reading.

Objectivity

Tests should be as objective as possible, that is, should use the same standard to measure all students. This does not mean that you can use only objective tests, but it does mean that when you use subjective tests, you need to establish objective ways to grade, and you have to decrease the difference in standards as much as possible. When you test language in ways that cannot be tested completely objectively, you need to set up concrete grading scales and train graders to use them properly. It is also useful to use more than one grader for each test.

For example, in testing writing, it is useful to make a scale, with separate sections for different aspects of writing, such as content, vocabulary, organization, and grammar, and a description of what would be expected at each level in each section. Such a scale helps the graders be as objective as possible about scoring compositions.

Practicality

Practicality is similar to simplicity. Tests should be as easy and practical as possible to administer, within the limitations of what you are trying to test. For example, if you are doing an interview test alone and need to videotape it, you may have trouble both operating the video camera and carrying out the interview. If you have only a short time to grade the test, and you have a large number of students, it is not practical to test the students by having them write long essays.

Economy

Unfortunately whatever you do, you have to work within a certain budget. You have to make and administer tests within your budget. If your test requires color copying, it will be more expensive. If you need to do recording, particularly video recording, it may be expensive. Tests that are be difficult to administer or grade will be more expensive, since the most expensive part is often labor. Thus, if the test requires many people to administer or grade it, it will be very expensive. It is always important to think of the cost of the test.

Scorability

Scorability is related to practicality and economy of the test. In some cases, particularly when testing writing or speaking, it is necessary to make relatively complicated scoring systems in order to make scoring more objective. However, if the procedure is complicated, it will cost more, it will take longer, and it will include more errors and/or subjective judgments. Therefore, it is necessary to consider the best way to score a test, to find a balance between a scoring system that is fair to the students and objective and one that is reasonably easy to administer.

Pretesting

It is important to pre-test the tests you plan to use. Even if you believe you have written a clear question that is easy to respond to and which has only one answer, you may well find that the question is ambiguous, has answers you had not considered, or is impossible to respond to. However, if you have already administered the test to your students, it is too late to revise the questions. If at all possible, you should pre-test the test on students similar to the ones who will take the final version of the test. If that is not possible, it is best to test it on non-native English speakers. If neither is possible, you can have native English speakers pre-test it, but there are obvious limitations to this procedure.

When pre-testing a test, you should ask the people who are pre-testing it to try to respond to the questions as if they were students taking the

test. Some problems may be obvious just from reading and thinking about the questions, but there are likely to be problems that will not be obvious without actually trying to answer the questions. Also, you can use the answers from the pre-test to the questions to see whether there are possible answers you had not considered and to try out your scoring criteria.

Conclusion

Making economical, practical, scorable tests which meet your goals and which are of an appropriate level of difficulty is an important part of English instruction. These factors should always be considered in making tests.

Discussion Questions

Look at a test that you have taken in the past (or a test that your teacher provides). Answer the following questions.

1. What appears to be the goal of the test, that is, what is it intended to test? Do you think it meets that goal?
2. Who is the intended audience of the test? Is it appropriate for that group?
3. Are the instructions clear? If not, why not?
4. Does the test appear to be practical, economical, simple, and scorable?
5. If it were your job to improve the test, how would you improve it? Why?

Chapter 7

Some Common Problems with Test Items

Key Words

grammar items
vocabulary items
pronunciation items
functions
reading
multiple choice questions
true/false questions

Introduction

There are certain problems that typically come up in test items or types of tests. Writing a perfect test item is difficult or impossible, and you cannot necessarily avoid all problems. However, being aware of some of the common problems will help you avoid them or compensate for them in other ways, or at least to take the issue into account in evaluating the results. In this chapter, we will discuss some of the common problems in test items and, in some cases, how they might be dealt with.

Item Types

Grammar Items

1. (an item intended to test the present perfect)

Write a sentence containing "since."

1)_____

In this case, the item may not elicit the present perfect. There are other uses of "since," and the testees may not produce the intended form. Another way of eliciting the present perfect is to give the testees a

sentence where it is necessary, along with the verb, and have them change the verb to the correct form.

Whenever you write an item that requires testees to produce something, there is a possibility that what they produce will not be what you intended. It is best to pretest the item to see what sorts of interpretations testees have of the item. If it is not possible to pretest an item on the type of people who will take the test, you should at least have some other people try to answer the item.

Vocabulary

Choose the word that has the same meaning as the given word.

1. catastrophe
   a. cataract
   b. conservation
   c. contiguous
   d. calamity

In this case the alternative that testees are asked to choose from are as difficult, if not more difficult than, the word that is being tested. The testees might know "catastrophe" but not "calamity" and thus would miss the item even though they knew the meaning of the word being tested. In the case of vocabulary items (as well as other types of multiple choice questions), it is important that the alternatives be easy to understand so that it is the head word, the content of the reading, etc., that is being tested—not the ability to understand the alternatives.

2. Use each word in a sentence that shows that you understand the meaning of the word:

   government

   industry

   negotiations

This test item really tests two things—the testees' knowledge of the meaning of these words and their ability to write sentences that display that knowledge. Even for a native English speaker, the latter could be quite difficult. While it is useful to have the testees deal with the word in

context, the problem with this approach is that the testees may understand the word perfectly well but not be able to make a sentence that shows its meaning.

Pronunciation

Which two words are not like the others? Consider the vowel sounds.

a. fail
b. male
c. land
d. reign
e. beak

This is intended to be a test of ability to pronounce, but studies have shown that there is no correlation between an ability to do well on such pencil-and-paper tests and the ability to actually pronounce English. Therefore, this is not a good pronunciation test.

Functions

Circle the letter of the function that each utterance serves.

Mike: Hey, give me back my coat!

A. command
B. complaint
C. accusation
D. request

In this case, there is more than one possible correct answer. Since the utterance is in the imperative form, it is probably a command though, depending on the context, it is possible that it is a request. However, it may also have the function of being an accusation (implying that the hearer took the coat without the speaker's permission) or a complaint (for taking the coat without permission). This item demonstrates two possible problems. First, when using multiple choice questions, it is very important to provide only one right answer. (Multiple choice questions are discussed in greater detail below.) It is necessary in writing multiple choice questions to try them out on students in advance, if possible, or, at least to have another person read the test and make judgments about

whether there are any questions that have more than one answer. Second, utterances can serve more than one function, and it may be difficult to tell which one is intended, especially if the utterance is not presented in context.

Reading

1. (After a passage on how Japan and the U.S. are different)

One way in which the United States and Japan are different is
a. Japan is a smaller country.
b. The U.S. has a more hetrogeneous population.
c. There are more people in the U.S.
d. Japan has a longer history.

Based on general knowledge, all of these answers are correct. It should be emphasized in the question itself or in the instructions for all of the questions that testees should answer the questions based on the information in the reading passage alone. Even so, this type of question might be difficult for lower level testees, because it might be difficult for the testees to distinguish between what they know and what they have read. Thus it might be best to avoid it when making tests for lower level tests.

2. Smallville is 1500 miles from Metropolis, and it takes three hours and fifteen minutes by plane, including a 45-minute stopover in Gotham City. One day, a plane left Smallville at 11:00 a.m. and had mechanical problems and had to stay an extra hour in Gotham City. When the plane took off again, it had to wait in line for thirty minutes. At what time did the plane arrive in Metropolis?

The problem with using this type of item as a reading test is that the testees may understand the content of the reading perfectly but not be able to answer the question, because they do not understand how to solve the puzzle. When writing reading items (or any other type of item), you need to carefully consider what skills or knowledge are necessary to answer the question. If a skill other than the one you want to test is necessary, that type of item is probably not a good item.

Format Types

Test items can use a variety of formats. There are also some common problems with writing the formats.

Multiple Choice Questions

Multiple choice questions are commonly used, because they are easy to grade, but they are more difficult to write than they appear. There are some common patterns used in writing multiple choice questions that should be avoided.

In writing multiple choice items, it is very important to write good correct and incorrect alternatives. The correct alternative in a multiple choice question should not come directly from the passage. It should be re-worded, so that the testees have to understand the meaning of the passage and they cannot just identify the repeated words. Incorrect alternatives should be based on a possible misunderstanding of the passage. They should be clearly incorrect, but not illogical. If they are illogical, the testees will eliminate them, even if they do not understand the passage. One problem in writing good multiple choice items is that it is often difficult to write three or four good alternatives that are both logical and clearly incorrect.

When writing multiple choice questions, you should avoid giving the testees clues to the correct answer in the way you write the alternatives. For example, there is a tendency for the correct answer to be much longer or much shorter than the other alternatives, so it is best if the alternatives are nearly the same length. Test item writers have a tendency to put the correct answer in the middle—that is, if there are four items, to make the correct item either B or C, or if there are five items, to make it C. You should make sure that the correct answers are not predominately in the middle of the alternatives. You should also avoid using the opposite of the correct alternative as an incorrect alternative, and you should avoid using two incorrect alternatives with very similar meanings.

When writing multiple choice questions, it is natural to make one of the distracters an opposite of the correct answer. Therefore, experienced test takers tend to assume that if there are two opposites among the

alternatives, one of them is probably correct. This is something that should be avoided in all types of multiple choice items, not just grammar items.

For a similar reason, you should avoid incorrect items that have the same or similar meanings. Experienced test takers realize that if there are two items with similar meanings, they must both be wrong, which eliminates them from being possible answers.

True/False Questions

True/false questions are commonly used, because they are easy to grade and relatively easy to write. However, you need to be aware of some potential problems. Consider the following example:

Read the passage and each statement that follows it. For each statement, indicate whether, according to the reading passage, it is true (T), false (F) or whether there is insufficient information to answer the question (I).

In a course I took in junior high school in Los Angeles, we discussed such problems as world hunger and homelessness, because those were problems that we thought were among the most important world issues at the time....

T F I 1. The students discussed important world issues.
T F I 2. The students in this class were 18 years old.
T F I 3. The students always discussed the topics they wanted to discuss.
T F I 4. The students discussed such problems as world hunger and homelessness.

One problem with true/false questions is that testees have a 50/50 chance of getting the correct answer, even if they do not know it. Adding "I" (insufficient information) makes it a little more difficult for testees to guess.

With the addition of the "I" alternative, the first question becomes difficult to answer. We know from reading the passage that the students in the class thought world hunger and homelessness were important world issues, but does the writer think so? Can it be understood that the writer

implies they are important issues, even though he/she does not say so directly? Arguments could be made for both points of view. When you write true/false questions, making an "insufficient information" option is useful, because it makes guessing less of a factor and decreases the possibility of getting a correct answer by chance. However, situations like this also arise—where there is a question about whether the writer implied something or not.

The answer to the second question is clearly "false"—but this comes from a combination of cultural knowledge and what is in the passage. The testees must know that students in American junior high schools are usually between the ages of 12 and 16, or they cannot answer the question correctly. Again, when writing a test, you need to be careful of questions that require information outside of the passage—information that some testees might lack—to be answered.

The third question uses the word "always." Since statements with words like "always" or "never" tend to be false (or at best lack sufficient information), they should be avoided.

The fourth question uses the same wording as the original passage. Testees might be able to answer the question by matching the words in the passage and the words in the question, without understanding the meaning. Therefore, the wording in a true/false item or a multiple choice item should be different from the wording of the passage being tested and, if possible, should be based on a potential misunderstanding of the meaning of the passage.

Like multiple choice items, false statements should be based on some possible misunderstanding of the passage. True statements should not be worded exactly like the passage, so that testees cannot just match the words with the question with the words in the passage.

Conclusion

In this chapter, you have seen various problems that can arise in making test items. As you write test items, you should keep these in mind and try to avoid them.

Application Exercise

1. Choose three of the items discussed in this chapter. Rewrite them to avoid the problems discussed. How have you corrected the problem? Does your rewritten item have another problem?

2. Bring in a test that you have taken in the past (or look at a test that the teacher provides for you). Do you see any of the problems discussed here in that test? If so, how could the items be rewritten to avoid the problem?

Section III
Writing Tests

Chapter 8
Testing Reading

Key Words

reading skills
choice of text
reading tasks

Introduction

Reading is probably the most common of the four skills to be tested, and it may seem to be the easiest of the skills to test. However, testing reading does have difficulties, and there are issues that anyone testing reading should be aware of. In this chapter, we will discuss the skills involved in reading and how they can be tested.

Skills Involved in Reading

Among the skills involved in reading are:

1. associating graphic symbols with sounds and words
2. understanding relationships between pieces of information in a sentence, including elements of sentence structure, negation, and embedding
3. deducing meanings of words from their roots and affixes
4. deducing meanings of words from the context

5. understanding presuppositions (e.g., "The driver did not stop for the stop sign" includes the presupposition that there was a stop sign.)
6. understanding relationships among parts of the text, signaled by such devices as lexical devices (synonyms, repetition, etc.), anaphoric reference (pronouns) and connectives (e.g., because, therefore)
7. understanding temporal and spatial relationships
8. understand relationships such as cause and effect; generalization and example; comparison; contrast; and opinion and support.
9. anticipating what will come next
10. identifying the main idea and supporting details
11. understanding figurative language and allegories
12. understanding inferences
13. skimming (getting an overall idea of the text)
14. scanning (locating specific information)
15. reading critically
16. varying reading strategies according to the type of text and the purpose of reading

While various lists have been made of the skills involved in reading, these are never clear cut. It is difficult to decide which skills are actually being used. It is also difficult to determine which skills are easiest, since what appears intuitively to be an easy skill might be one that gives readers a lot of difficulty and visa versa. However, from the point of view of making a test, lists of skills can be useful, since it encourages test makers to broaden their approach from just asking testees to find facts in the text.

Choosing a Text

The passage that is used to test reading can have an impact on the results. In this section, we will discuss some of the factors that need to be considered.

Variety

It is useful to use a variety of passages when testing reading. It is traditional to use prose to test reading skills, but there are many other

types of passages that can be used for testing. Some of these include schedules, newspaper articles, advertisements, and instructions.

Type of Text

The type of passage chosen should reflect the goals of the language teaching situation. If students are studying academic English, the passages should come from such sources. If they are studying English to use it in their daily lives, such texts as menus, leaflets, or letters might be useful.

Reading and Background Knowledge

One of the issues in testing reading that must be taken into account is that background knowledge plays an important part in reading comprehension. Briefly, when reading a passage for which they have some background knowledge, readers are able to comprehend it more easily and more completely and to make more accurate inferences from it than when reading a passage for which they have little background knowledge. Imagine, for example, reading an account of a game in a sport about which you know little or nothing, and compare that with reading an account of a game in a sport that you know a lot about. In the former case, you will probably understand little of the description of the game, and in the latter case, you will probably understand almost everything—even information that the writer does not state directly. If you are familiar with baseball, you know that a game with a score of 14-13 was an unusually high-scoring game, even if the writer does not say so. The reading passage chosen as the basis for the test must be one for which the testees are likely to have fairly similar levels of background knowledge. If not, testees who have background knowledge will have a great advantage over those who do not, whatever their reading proficiency.

In addition, the passage should not be about a subject that the testees are either very familiar with or very unfamiliar with. If students have too much background knowledge, they will probably be able to answer the questions without even understanding the test. If, in contrast, it is about a subject for which testees have almost no background knowledge, they may have a great deal of difficulty understanding the passage, and par-

ticularly making inferences, even if they are fairly good readers.

Difficulty

In choosing a passage, the test maker needs to take into account its difficulty. If the test is too difficult, few of the testees will do well; if it is too easy, too many of the testees will do well. However, the issue of difficulty is not a simple one. For subject areas for which they have good background knowledge, testees may be able to handle a more difficult passage than for subject areas for which they have little background knowledge.

Tasks to Test Reading

There are a wide variety of tasks that can be used to test reading ability. Since there are many more possible tasks than we can deal with in this chapter, we will present some of the more common and useful ones.

Testing Low Level Skills

There are some tasks that can be used to test very low level skills such as word recognition, sentence recognition, and word and sentence comprehension. These types of tasks have a place when testing students who are just beginning to read English, but they should not be used to test students beyond that basic level. It is useful to have testees do these tasks under time pressure, since they are intended to test skills that much be performed automatically in higher levels of reading.

Word recognition. Testees are presented with one word and a series of four or five words. They are instructed to mark the word that is the same as the first word.

Sentence recognition. This is similar to the previous task. Testees are presented with sentences instead of words and indicate which sentences are the same.

Word and picture matching. There are two variations on this task. The first is to give testees four similar pictures and one sentence. They are instructed to identify the picture that is described by the sentence. The other task is to look at one picture with four similar sentences and decide which of the sentences accurately describes the picture.

A variation on this is to give the testees, for example, ten sentences and have them choose five that accurately describe the picture.

Tasks for Middle and Higher Level Students

There are a wide variety of tasks that are possible to use to test reading comprehension. The following is a partial list.

True/false questions. Probably the most common type of task is true/false questions. They are particularly useful for progress tests, since they can be constructed relatively quickly and easily, and they are also easy to mark. The drawback of true/false tests, though, are that testees have a fifty/fifty chance of getting the correct answer, meaning that the facility value (a measure of the difficulty of the questions; see Chapter 17) is likely to be around 75 percent. This means that the test is relatively easy and may not discriminate well between high and low level students, unless there are a lot of questions.

There are at least two ways of dealing with this problem. One is to introduce a penalty for guessing. For example, two points could be awarded for each correct answer and one point deducted for each incorrect answer. Another way is to make a third alternative: the information is not in the reading. This type of question is sometimes difficult to write, though, because it may be difficult to make statements that appear close to the content of the reading passage without actually being correct. In other words, it is sometimes difficult to draw a line between what information can reasonably be inferred from the passage and what information is not there at all.

There are two types of true/false questions: questions that are independent of any passage and those that depend on the passage. In the case of questions that are independent, it is the testees' understanding of the language of the question itself that is being tested. In this case the content of the questions are about general knowledge all of the students could be expected to know, such as "Japan is smaller than the United States." Obviously it is necessary that all of the statements be ones that are within the knowledge of all of the testees, which is the drawback of this type of question.

For true/false tests that are dependent on the passage, testees read a passage and then answer questions directly related to the passage. These kinds of tests are frequently used for relatively elementary students, but they can also be written for intermediate and advanced students. Generally, in this case, the true/false test should test the understanding of the reading passage, not of the questions themselves. Therefore, it is important to make the questions themselves clear, concise, and easily understandable.

Multiple choice questions. Multiple choice tasks, like true/false questions, are easy to mark. They have the advantage over true false questions that there are more than two (or three) alternatives. Multiple choice questions can be made with four or possibly five alternatives.

One type of multiple choice item has a sentence or a few sentences, and testees choose from four words the one that fits in the context. For example:

We are moving to a city that has good schools, some nice parks, and safe streets. It is a good (environment, feeling, background, medium) to raise children.

This type of question can be used with testees of various levels. If testing vocabulary is the goal, then the context sentences should be easy, and the level of difficulty of the alternative words should be varied, depending on the presumed level of proficiency of the testees.

Another type of multiple choice question is one in which testees are given a sentence, and they are asked to decide which of four other sentences has the same meaning. This type of task tests understanding of grammatical structures.

Finally, multiple choice questions can be used to test the understanding of a reading passage. Testees can either be given a short reading passage with only one question or a longer passage with several questions.

It is often a temptation to concentrate on facts and figures. However, items should also test information that can be inferred from the reading passage or that require testees to put together information from more than one place in the passage.

BEST COPY AVAILABLE

Pre-testing tests is always important, but particularly so with multiple choice questions. It is very easy, for example, to make more than one correct alternative but not realize that a second alternative is also correct when considered from another point of view. Getting a fresh perspective on the items is important.

Short answer/completion questions. Various types of short answer questions can be used to test reading comprehension. These have the advantage that they require more than just recognition, but the problem with them is that they may also be testing production, along with reading. They are also more difficult to mark than true/false or multiple choice questions. Teachers will be presented with a range of answers, some correct, some incorrect, and some partially correct, and they have to decide how to deal with these.

A very common type of short answer question has a question to which the testees respond, using information from the passage. Another type gives a sentence stem, and the testees write in a word or words to complete the sentence, based on the information in the passage. Testees can be presented with a partially completed chart which they fill in with information from the passage.

Ordering tasks. Testees may read a passage and be given a series of statements covering the information in the passage to put in order. This is particularly useful for tests with a built-in sequencing, such as instructions or a narrative, but it can also be used to emphasize the development of ideas in the passage. An additional feature might be to include some statements with information which is not in the passage and instruct testees to indicate any such statements.

Charts. Charts can also be used to test reading comprehension. If you have a reading passage with two or more parallel sets of information (for example, dates and events), you make a chart of information from the passage. You leave some of the boxes in the chart blank, and students fill in the boxes with information from the passage. It is particularly important that it is clear what information is expected.

Main idea. An important part of reading is understanding the main idea. You can test the testees' understanding of the main idea of the entire

reading passage, a section of it, or even just a paragraph. There are various ways of testing the understanding of the main idea. You can just ask for the main idea, in either a short answer or multiple choice question. You can also ask students to choose or supply a title for the passage, section or paragraph.

Conclusion

Testing reading may seem easy at first glance, but the possible difficulties should not be overlooked. The choice or writing of the passage should be carefully considered so that they reflect the tasks for which the students will use English and so that background knowledge will not give some students advantages over others. In addition, the wording of the questions should be carefully considered and pretested.

Application Exercise

Decide on a situation in which you would give a reading test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a reading test. What type of passages would you choose or write? What task would you give? How are the decisions you made about the test tied to the situation you outlined above?

Chapter 9
Testing Listening

Key Words

phoneme discrimination
stress and intonation
sentences and dialogues
tasks using visual materials
tasks involving talks and lectures

Introduction

Testing listening involves a variety of skills. At the lowest level, it involves discrimination among sounds, discrimination among intonation and stress patterns, and comprehension of short and long listening texts. While the first two are part of listening, they are, of course, not sufficient. Testing of listening must also involve ability to understand the speakers' meaning and intention. In this chapter, we will discuss issues related to testing listening and methods of testing listening.

Testing Phoneme Discrimination

Sounds in the target language are sometimes difficult to discriminate if the sounds are not distinguished in the native language. There are several ways to test phoneme discrimination, that is, the ability to tell the difference between different sounds. One way is to have the testees look at a picture and listen to four words and decide which word is the object in the picture. The words chosen as alternatives should be close to the correct word. However, it is often difficult to find common enough words with similar sounds, and if unfamiliar words are used, they will not make good alternatives. Alternatively, the testees could be presented with four pictures and be asked to choose the picture that matches the word that

they hear. Another possibility is to give testees three words and ask them to indicate which two are the same. Finally, testees can listen to a spoken sentence and be asked to identify which one of four similar words were used in the sentence.

Items with full sentences have the drawback that testees can make use of not just phoneme discrimination but also knowledge of grammar and lexical items. If one of the words that the alternatives does not fit grammatically or semantically in the sentence, then testees who realize that have an advantage.

This type of discrimination item is one that can be used for diagnostic purposes to see whether students have particular problems with distinguishing between phonemes. However, it does not give the teacher any information about the testees' ability to comprehend spoken English.

Discriminating Stress and Intonation

The ability to recognize stress can be tested by having testees listen to a sentence that they also have in front of them. Testees are instructed to indicate the word that carries the main stress of the sentence. While recognizing stress patterns is useful in English, the problem with this type of test is that it lacks a context. Testees need to show that they can recognize the difference between "*John* is going today" and "John is going *today*," but they do not need to show that they understand that there is a difference in the meaning of the two sentence or what the difference is.

Ability to understand the meaning of difference in intonation can be tested by having the testees listen to a statement and choose from three interpretations of the statement. For example, testees might be given the statement "Vera is a wonderful musician" and be asked to decide whether the speaker is making a straightforward statement, a sarcastic statement, or a question. Since the context is neutral, however, it is sometimes difficult to avoid ambiguity. In real communication, listeners make use of their background knowledge, the context, etc., as well as the intonation to help them interpret the communicative meaning of an utterance.

Understanding Sentences and Dialogues

A teacher can also test the students' understanding of individual sentences and dialogues. In the simplest form, this type of item consists of a single sentence which testees listen to and four written statements from which they choose the one closest in meaning to the original spoken sentence. For example:

Spoken:
I had hoped to visit you while I was in New York.

Written:
A. I was in New York but did not visit you.
B. I will be in New York and hope to visit you.
C. I visited you in New York and hope to again.
D. I am in New York and would like to visit you.

Another type of item is one in which the testees listen to an utterance and choose from among four responses the most appropriate response. In that case, the testees are not being asked directly what the meaning of the utterance is. Rather they are being asked to show that they know what it means by showing that they recognize an appropriate response. This tests both the testees' listening ability and their knowledge of appropriate responses in natural conversation.

An example of this type of item is as follows.

Spoken:
Would you mind if I visited you next time I came to New York?

Written:
A. Yes, of course. I'd love to visit New York.
B. No, I don't really think that much of New York.
C. Yes, I would. You can come any time.
D. No, not at all. I'd really love to have you.

(At a slightly higher level, both the first statement and the responses can be spoken, but in that case, it might be better to have only three responses, since it would be difficult to keep all four responses in mind.)

In this example, the testees need to know that "Would you mind if I..." is a form used for asking permission, and that a positive response begins with "no (I don't mind)." Because this type of item requires two different types of information, there is a certain amount of controversy about it. Some theorists argue that it is not a good item type, because it requires these two types of knowledge. Testees could possibly understand the utterance perfectly well but not know how to respond to it. Also, since the utterances are presented in isolation and out of context, the situation is not realistic. However, this type of item can be useful if these limitations are kept in mind. It is a more communicative type of task than many other listening tasks, so it may have beneficial backwash effects, and it is relatively easy to administer.

Tasks Using Visual Materials

Matching and True/False Tasks

Some types of tasks make use of visual materials along with the spoken material to test listening. The simplest form of this task is to present testees with a picture or other visual (for example, a chart, graph, etc.) along with spoken true/false statements. Testees look at the visual and decide if the statements are true or false. An alternative is to present testees with a series of similar pictures and to have them match the pictures with spoken sentences describing them. Similarly, the testees can listen to a short dialogue, rather than just a statement, and decide which of the pictures matches the dialogue.

The advantage of using visual materials in this way is that they can be used to test listening alone without involving other skills very much (if the tester considers that an advantage)—though it is impossible to entirely eliminate the use of other skills. However, they may not reflect the sorts of listening done in the real world.

Map Tasks

Another way to use visual materials in testing listening is to use maps. One activity involves having testees listen to directions for how to get somewhere and follow along on the map. They respond by drawing their

route or indicating where they would be at the end of the directions. Another possibility is to have testees listen to a conversation referring to various locations on the map and having the testees identify the locations.

Drawing Tasks

Testees can also do drawing tasks according to instructions. For example, they can be given a simple line drawing and be asked to complete it according to certain instructions. The testees can be presented with a diagram of a room with the bed represented by a rectangular square and be asked to add a table, a bookcase, a door, etc., in certain locations in the room. It is important to keep such activities simple so that the drawing task itself does not demand too much of testees. The activity is more interesting if it can be done as part of a simple story rather than as a list of statements, though this may depend on the level of the testees. In addition, the students should have a chance to try out this type of activity before having to do it as part of a test.

Tasks Involving Talks and Lectures

For students who will be using English in schools where it is the medium of instruction, there will be situations where they need to listen to lectures or talks in English and take notes and/or answer questions on the lecture or talk. Therefore, listening tests can involve listening to formal or informal talks.

One way of using talks in listening tests is to have the testees listen to the talk and then fill in the blanks in a written summary of the talk. The words chosen from blanks should be ones that the testees cannot figure out from the context of the summary, without listening to the talk. However, they should also be ones that are related to the main idea of the talk, so that filling in the blanks does not require remembering small details of the talk.

Another way of doing this involves giving testees questions to answer as they listen to the talk. These questions can be short answer/completion, multiple choice, or true/false. The difficulty with using short

answer or completion questions is that they require the testees to both read and write while they are listening, something that can be difficult even for native speakers. Multiple choice questions may require a lot of reading, something that may also be a problem. True/false questions may be the best type for this type of task, since they require relatively little reading compared to multiple choice questions and relatively little writing compared to short answer/completion questions. Also, true/false questions can also have a "insufficient information" option, meaning that the information required to answer the question is not included in the talk. This decreases the amount of guessing and reduces the element of chance.

Another type of task that can be used is a chart that the testees fill in while listening to the talk. Answers in some of the blanks in the chart may be filled in for the testees. Testees should be given the chart in advance of hearing the talk, and be given time to familiarize themselves with it and make sure they understand what is expected of them.

Finally, testees can be allowed to take notes while listening to the talk, and then use the notes to answer questions after the talk is over. Depending on the length and complexity of the talk, testees might be given a list of the major topics included in the talk in order to help them in taking notes.

Conclusion

There are a number of ways to test listening, but, particularly when testees' listening proficiency gets more advanced, testing listening becomes more complicated. It becomes more difficult to separate listening from other skills, and combining skills can put great demands on the testee. In addition, some ways of testing listening do not reflect real-world listening tasks. In choosing tasks for listening, the teacher should be aware of these problems.

Application Exercise

Decide on a situation in which you would give a listening test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a listening test. What type of texts, if any, would you choose or write? What tasks would you give? What listening skills would you test? How are the decisions you made about the test tied to the situation you outlined above?

Chapter 10
Testing Writing

Key Words

components of writing
gap filling
form completion
making corrections
letter writing
essay writing
instructions
choice of topic
choice of tasks

Introduction

Testing each skill is uniquely difficult, but testing writing presents two particular problems. The first is making decisions about the matter of control, objectivity of the evaluation, and naturalness in the writing test. If you decide to test writing in a controlled way and in a way that can be graded with straightforward objectivity, you must do so in a way that does not necessarily reflect how writing is used by the students in the real world. If, on the other hand, you test writing in a way that would reflect how the students use writing in the real world, it is difficult to have control over the writing and to evaluate the student's work objectively.

The second major problem with testing writing is, if the test is done in a way that it cannot be graded objectively, it is necessary to develop a scale that allows it to be graded as objectively as possible. How this can be done is one of the great difficulties of testing writing. In this chapter, we will discuss these issues.

Components of Writing

The ability to write involves at least seven component skills. They are:

1. Grammatical ability. This is the ability to write English in grammatically correct sentences.
3. Lexical ability. The ability to choose words that are correct and used appropriately.
4. Mechanical ability. The ability to correctly use punctuation, spelling, capitalizaton, etc.
5. Stylistic skills. The ability to use sentences and paragraphs appropriately.
6. Organizational skills. The ability to organize written work according to the conventions of English, including the order and selection of material.
7. Judgements of appropriacy. The ability to make judgements about what is appropriate depending on the task, the purpose of the writing, and the audience.

Perhaps the most difficult—and most important—of these skills is the last. Native English speakers develop a sense of what is appropriate in different writing situations (though they may be taught to use specialized registers, such as academic English or business English). Registers of English range from very informal forms that use colloquialisms, slang, and jargon to standard English to more formal forms, such as the language used for business letters, legal documents, and academic papers. Writers must be aware of these differences and learn to follow the conventions of different situations. A writing test needs to take these skills into account.

Types of Writing Tasks

Gap Filling

One of the most controlled ways of testing writing is gap filling. Testees are presented with a text with blanks, and they fill in the blanks. This is a mixture of both reading and writing skills, which is sometimes a

problem, because it makes it difficult to decide what the scores really mean. However, with lower level students, it might be the only reasonable test of productive ability.

Form Completion

Another controlled way of testing writing is to have the testees fill out a form, for example, an application. The advantage of such a task is that it is at least somewhat communicative, but the disadvantage is that it does not require any connected discourse or any use of language greater than lexical knowledge and a small amount of grammar.

Making Corrections

In some situations, testees are presented with a short piece of writing which has deliberate grammar, punctuation and spelling errors, and they are asked to correct the errors. This task is related to one thing that people do when they write—editing—and it can be objectively corrected, but it does not represent the writing task as a whole.

Letter Writing

Letter writing is a common task for writing tests. The stimulus for the letter may be a situation that is explained in the instructions, a letter to which the testees are instructed to respond, information given in chart or graph form that is to be summarized in the testees' letter, pictures or drawings that give information about a situation the testees are expected to write a letter about, etc. In all of these possibilities, the tester must keep in mind that the situation must be as clear as possible for the testee, unless there is an intention to test reading and writing together. If the testee does not understand, for example, a letter that he/she is expected to respond to, it will be impossible to get a sample of writing to evaluate.

Essay Writing

Essay writing is probably one of the more common writing tasks, but it should be used carefully. If the future situation of the students will not include writing essays, the tester should carefully consider whether it is the best test of the students' writing ability.

Issues Related to Intermediate and Advanced Writing Tests

For beginners, it is often necessary to test writing indirectly, for example, by giving them very limited tasks that involve such activities as filling in blanks, which do not require them to write connected prose. However, once students get to the intermediate and advanced stages, their writing proficiency should be tested in direct ways, ways that allow students to demonstrate their ability to used connected discourse in appropriate ways, which is the way that they will use writing most in real life. The following is a list of desirable characteristics of such tests. For practical and other reasons, it may not be possible for a test to have all of these characteristics, but a teacher constructing a writing test should carefully consider the issues related to them.

Instructions

The instructions should be as short as possible, consistent with making the writing task and situation clear, so that the testee does not spend too much time on the instructions. Definitions should be provided in the instructions, if any are necessary.

Time and length. The instructions should also indicate the time allowed for the task and any instructions related to length in number of words. It may be useful to give the students a range of the number of words, so they have an idea of what is expected, though they should be encouraged not to spend too much time counting words.

Establishing a context and communicative purpose. The test should establish a clear purpose of communicating, especially by indicating the intended reader and giving a realistic context for the reader. In making the context realistic, the teacher should keep in mind, for example, that the testee would probably not use English in writing to someone who shares his/her native language, so instructions to write a letter to a family member would not be an appropriate use of context. Establishing the context should be done in as few words as possible, while still making the situation clear.

Choice of Topic(s)

Choosing the topic is a very important part of the writing task. If the testees find that the topic is something that they cannot respond to, they will not be able to show their writing ability. There are various approaches to choosing a topic and various issues to consider. There are not always clear answers about what type of topic is best, but the teacher should be aware of the issues involved.

Doability. The topic should be one that the testees are able to respond to. To determine this, it is necessary to pre-test the topic. A topic that sounds very reasonable in theory may not work well in reality.

Knowledge. The tester must consider the background knowledge that the students have about the topic. The topic should be one that the students have an approximately equal familiarity with. If students have different levels of knowledge about a topic, the ones with greater knowledge about it will have a clear advantage in writing about the topic. One approach is to choose a topic about which the testees are equally ignorant. However, this requires giving them enough information so that they can write, which may result in the testing of the reading skill as well as writing, unless the information can be presented visually. Otherwise, a topic must be chosen about which the students are approximately equally knowledgeable.

Motivation. The tester should consider the issue of motivation. Will the topic motivate students of the age, sex, field of study, background, etc., of the testees to write? It is often difficult to find a topic that will motivate all students equally. Some testers choose the strategy of choosing a subject that none of the students are likely to be motivated by. If none of the students are motivated, they will at least be on equal footing. The problem with this approach is that it may be difficult for students to do their best at showing their writing skill if they are not motivated by the topic.

Breadth. The topic needs to be broad enough that every testee can approach it from some angle. If the topic is too narrow, the testees have little flexibility in their approach to it and may not have an opportunity to show their writing proficiency.

Allowing students to choose topics. In some cases, students are allowed to choose from a list of topics. This raises difficulties in the reliability of the grading, so unless skill in choosing a topic is an ability being tested, this is not recommended.

Choice of Task(s)

Connected discourse. The task should require testees to write a piece of connected discourse. While there may be valid arguments for testing the writing of beginning students by having them just fill in blanks, once students are beyond the beginning stage, their writing proficiency should not be tested by having them translate from their native language or fill in blanks.

Realistic task. The task that is chosen should reflect the type of writing that the testee is required to do in the real world. If the test is for students who will be going to English-medium universities, an appropriate task would be having students write an essay on an academic topic. Tasks which the testee would not usually perform in English would not be appropriate.

Clarity. The testees should be presented with a clearly defined task that cannot easily be misinterpreted. Pre-testing helps insure that the instructions are clear and that the testees can carry out the task based on them.

Modes of discourse. The test tasks should involve a mode or modes of discourse that are appropriate to the actual writing needs of the students. If necessary, the testees should be given more than one task so that they can demonstrate their mastery of different modes of discourse. In fact, a recent trend is to evaluate students, where possible, on different types of material that they have produced over a long period of time, rather than over one piece of writing on a particular occasion.

The number of tasks. The test should involve more than one task, which will give an adequate sample of the testees' writing for evaluation. As mentioned above, different types of writing will also give a broader view of the students' writing skills.

Miscellaneous

Level of difficulty. The teacher should carefully consider the difficulty of the test. Like any other test, if a writing test presents a task that is too easy or too difficult, if the instructions are difficult to understand, etc., the responses that testees give will not reflect their true ability, either because the task is not challenging enough for their ability or because it is so difficult that they do not know how to respond. Pre-testing with a similar group is useful in determining the right level of difficulty.

Time allowed. The teacher should carefully consider the time allowed for the test. If insufficient time is allowed, the students do not have a chance to show what they can do, particularly on a test where the organization of a piece of writing is assessed. Students need enough time to produce a sufficient amount of writing to be assessed. However, there may be cases, such as writing an essay for an examination, where the task that the student needs to be able to do in the real world will have restrictions on time, so in some cases, it is appropriate to limit the time allowed.

Marking Tests of Writing

The marking of writing tests will always be at least somewhat subjective, but the use of descriptors for each level of the marking scheme can at least help make the marking consistent. One possibility is to make a marking scheme for the overall quality of the writing, but the problem is that, for example, the grammar can be good but the organization poor. It is perhaps more useful to have different sets of descriptors for each aspect of writing that you want to consider. You might want to have descriptors for grammatical correctness, use of vocabulary, content, organization, and mechanics. These categories might be weighted differently, depending on what you want to emphasize.

Conclusion

Teachers can use a number of tasks to test students' writing, ranging from gap-filling exercises to essays, but the tasks should be chosen

according to the students' level and their future uses for English. Issues such as the choice of topics and the choice of tasks must be carefully considered.

Application Exercise

Decide on a situation in which you would give a writing test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a writing test. What type of task would you give? How would you assess the writing? How are the decisions you made about the test tied to the situation you outlined above?

Sample task:

Write a letter to the Ms. Linda McNeal, head of the Superior Language School at 435 Gloucester Road, Boston, Massachusetts 01774 USA. As for information about courses offered at the school, the possibility of financial assistance, and accommodation near the school.

Chapter 11
Testing Speaking

Key Words

reading aloud
conversational exchanges
visual materials
role plays
group/pair activities
assessment

Introduction

Speaking is a difficult skill to test. In this chapter, we will discuss what makes speaking difficult to test, and how it can be tested, in spite of these difficulties.

Difficulties of Testing Speaking

Speaking is probably the most difficult skill to test. It involves a combination of skills that may have no correlation with each other, and that do not lend themselves well to objective testing. There are not yet good answers to questions about the criteria for testing these skills and the weighting of these factors. A speaker can produce all the right sounds but not make any sense, or have great difficulties with phonology and grammar and yet be able to get the message across clearly. Also, success in speaking depends to a great extent on the listener. Comprehension of spoken material depends, among other factors, on the degree to which the listener is familiar with the speaker's accent and the degree to which they share background knowledge, and so what is a problem for one listener may not be a problem for another listener.

Another difficulty is separating the listening skill from the speaking

skill. In most normal situations, there is an interchange between listening and speaking, and speaking appropriately depends, in part, on comprehending spoken input. This necessarily becomes a factor in the testing of speaking, and it is difficult to know whether you are testing purely speaking or speaking and listening together. (However, if those administering the test are taking a communicative approach, they may not consider this a problem.)

Testing speaking is also a particular problem when it is necessary to test large numbers of students. In some situations, it is necessary to test thousands of students, and even if each student speaks for only a few minutes, this becomes a huge job.

In spite of the difficulties inherent in testing speaking, a speaking test can be a source of beneficial backwash. If speaking is tested—unless it is tested in a mechanical way, such as reading aloud—this encourages the teaching of speaking in classes.

Types of Speaking Tests

Reading Aloud

One way that reading is tested is by having the testees read aloud. This is not generally a good way to test speaking. Its backwash effect is likely to be harmful, and it is not a skill that is used much outside of the classroom. However, it is a way to test pronunciation separately from the content of speech. If it is necessary to use this method of testing, the test should at least make use of a situation where the student might actually be reading aloud, such as reading instructions or parts of a letter to another person.

Conversational Exchanges

Another simple type of test is a test in which students are given a particular situation and instructed to respond in a certain way. These tests are usually highly structured and require only a limited response, not connected discourse.

Type 1. Students are asked to construct a sentence following a certain pattern using the information that they are given. They are usually given

two or three examples first. For example:

1. Jane has been working at the same job for ten years. She is tired of her job and would like to do something more interesting. (*She wishes she could find a more interesting job.*)

2. It has been raining since yesterday. I want to go on a picnic this afternoon. (*I wish it would stop raining.*)

Type 2. In another type of conversational exchange test, a situation is given but no guidance is given as to how to respond, and the student can respond freely. For example:

A friend of yours has been tired a lot recently. What do you say to her? (*Maybe you should be getting more rest.*, etc.)

You have been served coffee in a restaurant but you have not been given any cream or sugar, and you would like some. What do you say to the server? (*Could I have some cream and sugar.*, etc.)

Type 3. In another type of test, the student is given a stimulus sentence and can respond in any appropriate way. This type of test usually depends on conventional responses, such as responses to requests, invitations, etc.

Could I borrow your dictionary?
(*Sure, go ahead., Sorry, I'm using it.*, etc.)

Hey, how are you doing?
(*Just fine, how are you?* etc.)

This may also be done as part of a longer dialogue. For example, the test may include a pre-recorded tape of one side of a conversation, with pauses where the student is expected to respond to the utterances on the tape. The student listens to the tape and, where there are pauses for the testee's responses, the testee produces that he/she thinks is appropriate at that point in the coversation. The problem with this sort of test is that the dialogue on the tape keeps going along, no matter what the testee says, whether the next line on the tape is appropriate or not. In addition, it is testing listening at least as much as it is testing speaking. (This is not

necessarily a problem, as long as those using the results of the test are aware of it.) However, this type of test does make it possible to test large numbers of people at the same time in a way that is at least somewhat communicative.

Testing Speaking Using Visual Material

Pictures, maps, diagrams, and other types of visual material can be used to test speaking without requiring the testee to comprehend written or spoken material. Through careful selection of the material, the tester can control the vocabulary and, to some extent, the grammatical structures required.

Various types of material are appropriate for this type of test, depending on the language that the tester wants to elicit. One common type of stimulus material is a series of pictures which tell a story, often with a few sentences of introduction to get the testee started. This requires the testee to put together a coherent narrative. A variation on this is to give the pictures in random order of the narrative to a group of testees. The students describe their pictures and decide on their sequence without showing them to each other, and then put them down in the order that they have decided on. They then have an opportunity to reorder the pictures if they feel it is necessary.

Another type of test using a visual stimulus is one in which two testees are given pictures that are similar but have several differences. Without seeing each other's pictures, they describe their own pictures and find the differences. When testing two students together like this, it is important that their ability be somewhat similar, so that the lower-ability student does not bring down the score of his or her partner.

In addition to using pictures, it is possible to use charts, graphs, or other ways of visually organizing information, and the testee has to describe them. One possible problem with this type of visual stimuli is that the material has to be something that all the testees can interpret equally well, since if a testee has difficulty understanding the visual information, that will influence the way he/she is evaluated.

Oral Interviews

Oral interviews are testing situations in which the tester and the testee carry on a conversation. The tester generally has a list of questions to ask the testee, and someone—either the interviewer or another person, but preferably another person—assesses the language proficiency of the testee. The advantage of an interview is that it attempts to approximate a conversation situation, but the knowledge that it is a test and the constraints of the roles in the interview make it difficult to have a real, natural conversation.

The interview protocol needs to be carefully considered. It might be best to start with yes/no questions or factual questions that are easily answered to put the interviewee at ease, but this type of question does not give the student much scope for responding, so this stage should not last too long. The questions should be interesting, easy to respond to, and give the testee some range in answering.

When doing an interview test, it is very important that the interviewer be well trained in order to avoid problems and so that the interviewer is consistent from one interview to another to help make the evaluation of the testee more reliable. The interviewer, for example, should avoid talking too much (though giving some personal information may help break the ice and put the testee at ease), should speak clearly, should not interrupt the interviewee to give answers, should maintain eye contact with the interviewee, and should not correct the interviewee during the interview.

Role Play Test

Another type of test is a role play. In a role play the testee and a confederate or another testee are given information on which to base a role play, and the testees are evaluated on their ability to carry out the task in the role play. For example, the role play might be getting information about course requirements. Role plays require the testee to use various functions that he/she might need in real communication.

Group or Pair Activities

To avoid the types of problems associated with oral interviews,

sometimes students are put together in pairs or groups that are given some task, such as the ones mentioned above in the section on using visual materials. They might also be given a quiz to work on together, a puzzle to work out, or a task. The point of these is not to find the "right" answer but to stimulate speech for the tester to evaluate. This is intended to be a less stressful situation than an interview test, but it also has the problem of lack of control. How the activity goes depends to a great extent on the mix of testees that are put together. If their level of English proficiency is widely different, or if one testee is particularly dominant, this will influence the interaction. A well-trained interviewer can provide similar environments from one interview to another, but this is not possible in testing a group of testees in a group activity. In addition, successful group discussions depend on factors other than speaking, including active listening and eliciting the opinions of all of the group members.

Assessment

One of the great difficulties in testing speaking is, of course, the assessment. It is necessary to develop a system of assessment that can be applied as objectively as possible, though it is probably never possible to avoid some subjectivity in assessment. The scale can be one general scale for overall speaking ability, or it can be divided between several aspects of the skill of speaking, such as pronunciation, grammar, organization, etc. The scale also depends on the speaking task that is used for the test. A test that uses public speaking as the task would be different from one that uses a group discussion.

If possible, the speaking task should be recorded and the scoring done from the tape. In addition, the marking should be done by more than one person and the reliability of the raters checked. If the task is an interview, the interviewer should not be required to score the test while conducting the interview, if this is avoidable.

Among the aspects of speaking that might be considered in the assessment scale are grammar, pronunciation, fluency, organization, content and vocabulary.

The band descriptions for a general scale might be as follows. The number indicates the level, and it is followed by a description of the characteristics of a speaker at that level.

7 Spoken communication is fluent, appropriate, and grammatically correct, with few if any errors.

6 Communication is generally fluent and grammatically correct with only occasional errors in grammar or pronunciation.

5 Student produces numerous grammatical errors and hesitations, but these do not interfere greatly with communication. Utterances are long and connected.

4 Student produces numerous grammatical errors and hesitations, and these occasionally interfere with communication. Utterances are short and connected.

3 Student's communication is limited to short utterances and depends in part on previously memorized conversational elements. Difficulty dealing with unpredictable elements. Many hesitations and grammatical errors. Communication only possible with sympathetic interlocutor.

2 Communication limited to short utterances, almost entirely memorized conversational elements. Unable to deal with unpredictable elements.

1 No communication possible.

Conclusion

Though speaking is a particularly difficult skill to assess, there are methods that can be employed to create situations that elicit speech and methods of assessing the testees' speech that are reasonably reliable. Testing speech is important for its backwash effect, even if the method of testing and of assessment are not as perfect as they might be.

Application Exercise

Decide on a situation in which you would give a speaking test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a speaking test. What type of test would you give (i.e., interview, group activity, etc.)? What task would you give? How would you assess the speaking? How are the decisions you made about the test tied to the situation you outlined above?

Chapter 12
Testing Grammar

Key Words

multiple choice
error recognition
word/sentence order
completion items
transformation items
word changing items
sentence combining exercises

Introduction

The testing of grammar is one of the mainstays of language testing. Such tests usually evaluate the ability to either recognize or produce correct grammar and usage, but they do not test the ability to use the language to express meaning. However, it can be argued that a basic knowledge of grammar underlies the ability to use language to express meaning, and so grammar tests do have a part to play in language programs. In this chapter, we will discuss how it can be tested.

Types of Tests

Multiple Choice Tests

Probably the most common way of testing grammatical knowledge is the multiple choice test. These tests have the advantage of being easy to grade and being able to cover a lot of grammatical points quickly. However, they test only recognition, not production.

The most common type of multiple choice grammatical item is one in which the test maker gives the testee a sentence with a blank and four or five choices of a word or phrase which completes the sentence correctly.

For example,

Because my mother was sick, I _____ to go home last week.
a) had
b) have
c) has
d) haven't

To give slightly more context, this type of question sometimes makes use of a short dialogue, with one person saying something and the other person responding.

A way of testing short answers and responses is to give the testees an utterance, and have them decide which of four or five utterances is an appropriate response. This can be either a test of comprehension or a test of grammar. For example,

"I think that tuition is much too high here."
a) I do so.
b) Do I so.
c) I so do.
d) So do I.

Another way to test grammatical knowledge using multiple choice items is to give testees a sentence and ask them to choose which of four or five alternatives has the same meaning.

"The school should have expelled him."
a) The school did not expel him, which was wrong.
b) The school expelled him, because it was necessary.
c) The school might have expelled him, if it had known.
d) The school will probably expel him in the near future.

Again this is a test of reading comprehension as well as grammar, but in order to understand the meaning of the sentence, the reader does have to understand the grammar.

One way to give more context and make the language more natural is to give the items in the form of a short reading text rather than individual sentences. This gives the testees more context and, if the passage is

chosen carefully, it is also much more interesting than reading individual, uncontextualized sentences. However, it may be more difficult to test a range of grammatical points, since the grammatical points are restricted by the content of the text.

A variation on this idea is to use a piece of prose written by a nonnative English speaker. This is particularly useful for making a grammar test for testees who are all of the same language group, since the errors made by the writer can be used as distracters.

Error Recognition

Error recognition items are also useful for testing grammar. An error recognition item is one in which the testee is given a sentence with an error. Four words or phrases in the sentence marked with letters, and the testee needs to decide which of the words or phrases has the error. For example,

Most of students (A) believe (B) that they should be getting (C) better grades than they are. (D)

The teacher may also mix in some sentences that have no errors, where students are required to indicate that there is no error. In addition, the students might be required to correct the error. Errors from students' actual writing are a good source of ideas for this type of exercise.

Items to Test Knowledge of Word/Sentence Order

Other types of items can be used to test testees' knowledge of word order. The traditional way is to present the testee with four alternative word orders. For example,

I wonder how she knows
a) how it costs much.
b) how much it costs.
c) it costs how much.
d) it how much costs.

Another possibility is to give testees the four words and ask them to put the words in order. For example,

I wonder how she knows
A. how B. it C. much D. costs ☐☐☐☐

This can also be done in a way that actually requires the writer to do some writing. For example,

I wonder how she knows _____.
how / it /much / costs

Understanding of appropriate sentence order can also be tested in a similar way by giving testees several sentences and asking them to put the sentences in order. This type of test tests knowledge of references, cohesive devices (words and phrases that indicate the relationships between different pieces of information), etc.

Completion Items

Completion items are items in which the testees are asked to fill in blanks in sentences. For example,

Give the book to _____ woman in the blue dress.

For the purpose of a grammar test, the words that fit in the blanks should be function words, such as articles and prepositions. (Completion items intended to test reading ability or vocabulary knowledge, in contrast, use content words.) The advantage of completion items is that they test production, not just recognition. The disadvantage is that they need to be marked by hand and there will be some cases where the marker needs to make judgments about whether a response is correct. It is not always easy to write items for which there is only one possible answer. Using a piece of continuous prose rather than disconnected sentences is one way of cutting down on possible different interpretations of what goes into a particular blank, but one usually cannot entirely eliminate the possibility of alternate correct answers.

Also, it is possible to require a phrase instead of a word in each blank. However, while this method presents a more realistic situation, it does

become more difficult to mark, because there are more possible answers for each blank. While it may not be realistic for large-scale testing situations, it is something that is useful for classroom teachers who want to help their students develop an ability to produce appropriate grammatical forms in context.

Transformation Items

Another type of grammar item makes use of transformations. In this type of item, testees are given a sentence and the first few words of another sentence and instructed to change the original sentence without changing the meaning. For example,

1. Jim hasn't been home in a long time.
It's been a long time _____.

2. I don't need to go to the grocery store this week.
It isn't _____.

3. It is difficult to study when it is so noisy.
Studying _____.

There are variations on this type of item in which the word which starts the transformed sentence is underlined, or the testee is given one word to use in the new sentence. For example,

I don't need to go to the grocery store this week. (necessary)

Again, this type of test is difficult to grade because the teacher has to be aware of the variety of possible answers. Another problem is that it does not in any way test the testees' knowledge of when each of the possible transformations would be most appropriate. For example, the testee might be perfectly able to transform an active sentence to a passive sentence but not know when to use passive rather than active. However, it is still sometimes a useful test of grammatical knowledge.

Word Changing Items

Another type of item is one in which the testees are given a sentence and a word that they need to fit into the sentence by changing the form of the word. For example,

1. I have never _____ to Australia. (be)
2. I will be with you _____. (moment)

This type of grammar test item tests students' knowledge of different word forms and how they are used in sentences.

Sentence Combining Exercises

Sentence combining exercises can play a part in testing grammar as well as its more traditional use as part of composition testing and training. For example, testees might be instructed to combine the following sentences using a relative pronoun.

I met a man.
The man went to the same high school I did.

I met a man who went to the same high school I did.

Conclusion

While the testing of grammatical knowledge is limited—it does not necessarily indicate whether the testee can use the grammatical knowledge in a communicative situation—it is sometimes necessary and useful. When considering the testing of grammar, the teacher has to make decisions about such factors as ease of marking, the degree of control, and the degree of authenticity.

Application Exercise

Decide on a situation in which you would give a grammar test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a grammar test. What type of items would you write? Would you write items that require production or recognition or both? Would you prefer items that are easy to grade or ones that are more realistic? How are the decisions you made about the test tied to the situation you outlined above?

Chapter 13

Testing Vocabulary

Key Words

sources of words
synonym/definition items
picture items
sentence items
context items
set items
matching items
completion items
word formation items
guessing meaning from context

Introduction

A good knowledge of English vocabulary is important for anyone who wants to use the language, so knowledge of vocabulary is often tested. There are several issues related to the testing of vocabulary that a test maker should be aware of. In this chapter, we will discuss some of these issues and also techniques for testing vocabulary.

Types of Vocabulary Knowledge

Vocabulary knowledge can be divided into four types. The first is active speaking vocabulary, that is, words that the speaker is able to use in speaking. The second is passive listening vocabulary, which is words that the listener recognizes but cannot necessarily produce when speaking. The third type is passive reading vocabulary, which refers to words that a reader recognizes but would not necessarily be able to produce. Finally, there is active writing vocabulary, which is words that a writer is

able to use in writing. This typology can be expressed in the following chart:

	Oral	Graphic
Active	active speaking	active writing
Passive	passive listening	passive reading

A test may test one or all of these types of vocabulary, but the test maker should be aware of the differences among these types and which is being tested. Most traditional vocabulary tests test only passive reading vocabulary, since they are paper and pencil tests and rely on reading. If the test maker intends to test only passive reading vocabulary, that is fine, but there should be an awareness that that is what is being tested. If test specifications or course goals require other types of vocabulary to be tested, then other methods need to be used to test other types of vocabulary knowledge. Once the decision has been made to test other types of vocabulary knowledge, there must also be a decision about how to weight the four types of items. If the type of English education students have been exposed to emphasized reading, then it is appropriate to give passive reading vocabulary the greatest weight.

Sources of Words to Test

The test maker's job is easiest if all of the students have been using the same syllabus or course of study. In that case, sources include the syllabus, the textbook (with emphasis on words that occur in natural speech or writing), students' reading materials, and lexical errors from students' written work or incorrect answers on a cloze test.

If the testees have not all been using the same syllabus—for example, in the case of a proficiency test used to judge testees' suitability for a certain task—the problem of choosing vocabulary becomes more complicated. In that case, it is probably most appropriate to draw on the vocabulary of the task for which the testees are being tested. If the test is being used to choose students to go to an English-medium university,

then the items for the test can be chosen from academic vocabulary.

In any case, when testing vocabulary, the grammatical structures used should be easy to understand. If the grammatical structures are difficult to understand, the test is testing both the ability to understand the grammatical structure and knowledge of vocabulary. Therefore, a testee might miss an item due to a lack of understanding of the grammatical structure, not due to lack of vocabulary knowledge. (Conversely, when testing grammatical knowledge, difficult vocabulary should be avoided to avoid testing vocabulary rather than grammatical knowledge.)

Multiple Choice Tasks

Multiple choice tasks, in general, were dealt with in the chapter on common problems in writing test items. However, there are some aspects that are specific to vocabulary tests. Research indicates that the best distracters for vocabulary items are either 1) words that have a similar meaning to the correct word but which are inappropriate in context, or 2) words that look similar to the correct word but which do not fit in the context.

When choosing the four or five alternatives, care should be taken to choose words that are of a similar level of difficulty. If the correct alternative is much more difficult than any of the other alternatives, the testee might choose it not because he/she knows that it is correct but because all of the other, easier alternatives can be eliminated.

Alternatives should be the same part of speech as the word in the stem. If one is a different part of speech, that can easily be eliminated as a possible answer.

Synonym/Definition Items

The simplest multiple-choice item has a single word in the stem, with four or five alternatives. Testees choose from among the alternatives the word or the definition that is the closest in meaning to the word in the stem. Alternatively, the definition might be the stem, and the testees choose from four or five words that one that fits the definition. Obviously the definition needs to be simply and clearly written.

occupation
a. job b. interest c. office d. thought

Picture Items

The stem in the previous type of item can be replaced with a picture. The testees choose which of the four or five alternatives matches the picture. Obviously it should be clear what is being depicted in the picture, so that the test is of the meaning of the word, not the testee's ability to interpret the picture. This type of test is appropriate for elementary students.

Sentence Items

Another option is to test the vocabulary item by putting it in a sentence and having the testees choose the alternative that has the same meaning as a word in the sentence that is underlined, in italics, in bold type, etc. This is a preferred type of multiple-choice vocabulary item, since the problem item appears in context and the context helps give the vocabulary word a specific meaning. However, in this case, the testee should not be able to surmise the meaning of the word from the context, unless the intention of the item is to test that skill.

The climate in Southern California is pleasant.
a. weather b. scenery c. life d. food

Context Items

Another type of multiple choice item is one that has a sentence with a blank in it, and the testees choose which of the alternatives fit in the sentence. These items are difficult to write, because it is difficult to provide the right amount of context. If not enough context is provided, the question becomes difficult to answer; if too much context is provided, too much information, grammatical and semantic, is provided, and students use this, rather than their knowledge of lexical meaning, to answer the questions.

Sometimes such questions involve a grammatical component. For example,

Would you please _____ the proposal.
a. consider b. think c. make d. give

In this case, "think" (about) is similar to the correct answer, but it does not fit grammatically. While a few of these types of items might be acceptable, too many of them may put as much emphasis on grammar as on vocabulary.

Sometimes vocabulary items test collocations. For example,

I _____ my watch to see what time it was.
a. looked at b. saw c. watched d. gazed at

In this case, what is being tested is the knowledge of the collocation "look at my watch." Again, knowing collocations is part of knowing vocabulary, but the test should not be dominated by such items.

Set Items

Another type of item is one where testees are presented with a set of words. In the case of a recognition type of item, the testees decide which word in the set does not fit with the others; in the case of production type items, the testee indicates what topic ties all of the words together. For example,

Circle the word that does not fit.
dollar
yen
pound
money
(answer = money)

Write down the category that each group of words is related to.
bedroom
living room
kitchen
dining room
(category = rooms in a house)

The difficulty of these items is that they may be interpreted in different ways. In some cases, arguments could be made for excluding different words, depending on the way that the student thinks about them. Also, a student may understand all of the words but not be able to understand what the test maker is trying to get at.

Matching Items

Another type of item is one in which the testee is presented with a short passage or several unconnected sentences with blanks. The testee the word that fits in each blank chooses from among a list of words that are given (usually more words than blanks). In these items, it is preferable that all the words in a set be the same part of speech, the same tense in the case of verbs, etc., if it is important to test vocabulary meaning only, and not grammar. The purpose of having extra words is so that the testees cannot answer the last item by having eliminated all other possibilities. On the other hand, the more items there are, the more likelihood that two items could conceivably fit in the same blank.

A variation of this type of item is one in which the testee is given a reading passage and a list of words and is instructed to find synonyms in the passage for each word in the list.

Completion Items

Completion items are similar to those described above, except that testees are not given words from which to choose. They must supply the words for the blanks. The problem with this type of item is that it is very difficult to write items for which there is one and only one answer, which means that it is difficult to score the test objectively.

Word Formation Items

Another type of item tests not the testees' knowledge of the meaning of words but their knowledge of word forms. In this case, testees are given a sentence or paragraph, and they are instructed to fill in blanks with the correct forms of the words that they are given. (A variation on

this is to write multiple choice items, in which the stem is a sentence with a blank, and testees choose from among four forms of the same word.) For example,

Write the correct form of the word in capital letters in the blank.

HONEST
Darren says he didn't cheat, but I _____ don't know what to think.

Guessing Meaning from Context

Another vocabulary-related skill to test is the ability to guess the meaning of an unknown item from context. In testing this skill, students are usually presented with a word that they would be unlikely to know and are asked to figure out what it means. Among the clues from the context are:

1) a synonym in another clause, for example, "Your explanation doesn't need to be so *convoluted*. Such complex explanations just confuse people."
2) an antonym or contrasted word in another clause, for example, "May was *indignant* about the problem, and I had to calm her down."
3) a cause or effect of the word in question, for example, "Pollution from that factory is *contaminating* the local farmers' fields."
4) an illustration of the word in question, for example, "Jeff is so *parsimonious* that he won't spend a penny if he doesn't have to."
5) the use of an object, for example, "I used a *pitter* to remove the seeds from the cherries."

Testing this skill can be done as a multiple choice item or by requiring the testees to supply a definition. The problem with the latter type of item is that it is very difficult to evaluate the answers, because some are likely to be almost correct but not exactly. It is necessary that the meaning of the target word be very clear from the context, and for this type of item, pre-testing is particularly important for that reason.

BEST COPY AVAILABLE

Conclusion

Knowledge of vocabulary is important to language use, and it is useful to be able to test from various points of view—knowledge of word meanings, knowledge of word forms, and knowledge of how to surmise the meanings of unknown words from the context.

Application Exercise

Decide on a situation in which you would give a vocabulary test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a vocabulary test. How would you choose the vocabulary items for the test? What types of test items would you use? Would you test only the testees' knowledge of words, or would you include their knowledge of word forms and how to infer the meanings of unknown words from context? How are the decisions you made about the test tied to the situation you outlined above?

Chapter 14

Testing Communicative Competence

Key Words

communicative approach to language teaching
tasks
evaluation
information gap
role play
letter writing
summarizing

Introduction

Testing language has traditionally taken the form of testing knowledge about language, usually the testing of knowledge of vocabulary and grammar. However, there is much more to being able to use language than knowledge about it. Dell Hymes proposed the concept of communicative competence. He argued that a speaker who is linguistically competent can be able to produce grammatical sentences, but unless the speaker is also communicatively competent, the utterances may be completely inappropriate. In communicative competence, he included not only the ability to form correct sentences but to use them at appropriate times. Since Hymes proposed this idea in the early 1970s, it has been expanded considerably, and various types of competencies have been proposed. However, the basic idea of communicative competence remains the ability to use language appropriately, both receptively and productively, in real situations. In this chapter we will discuss ways that language can be tested communicatively.

The Communicative Approach to Testing

What Communicative Language Tests Measure

Communicative language tests are intended to be a measure of how well the testees are able to use language in real life situations. In testing productive skills, emphasis is placed on appropriateness rather than on ability to form grammatically correct sentences. In testing receptive skills, emphasis is placed on understanding the communicative intent of the speaker or writer rather than on picking out specific details. And, in fact, the two are often combined in communicative testing, so that the testee must both comprehend and respond in real time. In real life, the different skills are not often used entirely in isolation. Students in a class may listen to a lecture, but they may later need to use information from the lecture in a paper. In taking part in a group discussion, they need to use both listening and speaking skills. Even reading a book for pleasure may be followed by recommending it to a friend and telling the friend why you liked it.

The "communicativeness" of a test might be seen as being on a continuum. Few tests are completely communicative; many tests have some element of communicativeness. For example, a test in which testees listen to an utterance on a tape and then choose from among three choices the most appropriate response is more communicative than one in which the testees answer a question about the meaning of the utterance. However, it is less communicative than one in which the testees are face-to-face with the interlocutor (rather than listening to a tape) and are required to produce an appropriate response.

Tasks

Communicative tests are often context-specific. A test for testees who are going to British universities as students would be very different from one for testees who are going to their company's branch office in the United States. If at all possible, a communicative language test should be based on a description of the language that the testees need to be able to use. Though communicative testing is not limited to English for Specific Purposes situations, the test should reflect the communicative situation

in which the testees are likely to find themselves. In cases where the testees do not have a specific purpose, the language that they are tested on can be directed toward general social situations where they might be in a position to use English.

This basic assumption influences the tasks chosen to test language in communicative situations. A communicative test of listening, then, would test not whether the testee could understand what the utterance, "Would you mind putting the groceries away before you leave" means, but place it in a context and see if the testee can respond appropriately to it.

If students are going to be tested over communicative tasks in an achievement test situation, it is necessary that they be prepared for that kind of test—that is, that the course material should cover the sorts of tasks they are being asked to perform. For example, you cannot expect testees to correctly perform such functions as requests and apologies appropriately and evaluate them on this if they have been studying grammar apart from a communicative context. Similarly, if they have not been studying writing business letters, you cannot expect them to write a business letter for a test.

Tests intended to test communicative language are judged, then, more on the extent to which they simulate real life communicative situations than on how reliable the results are. In fact, there is an almost inevitable loss of reliability as a result of the loss of control in a communicative testing situation. If, for example, a test is intended to test the ability to participate in a group discussion for students who are going to a British university, it is impossible to control what the other participants in the discussion will say, so not every testee will be observed in the same situation, which would be ideal for test reliability. However, it would be almost impossible to accurately test this ability in a completely controlled situation, so according to the basic assumptions of communicative language testing, this lack of control is compensated for by the realism of the situation.

Evaluation

There is necessarily a subjective element to the evaluation of communicative tests. Real life situations do not always have objectively right or

wrong answers, and so band scales need to be developed to evaluate the results. Each band has a description of the quality (and sometimes quantity) of the receptive or productive performance of the testee. The following band scale may be used to evaluate a test that evaluates communicative competence in a group discussion situation.

4 Contributions to the group discussion are appropriate and substantive, although student does not dominate the discussion. Student elicits opinions of other group members.

3 Student contributes to the group appropriately and substantively to the discussion but may either dominate the conversation or fail to elicit contributions from other group members.

2 Student does contribute to the discussion, but some of the contributions may not be appropriate or substantive.

1 Student only rarely contributes to the discussion, or contributions are inappropriate or non-substantive.

0 Student makes no contributions to the discussion.

Examples of Communicative Test Tasks

Speaking/Listening

Information gap. An information gap activity is one in which two or more testees work together, though it is possible for a confederate of the examiner rather than a testee to take one of the parts. Each testee is given certain information but also lacks some necessary information. The task requires the testees to ask for and give information. The task should provide a context in which it is logical for the testees to be sharing information. The following is an example of an information gap activity.

Student A

You are planning to buy a tape recorder. You don't want to spend more than about 80 pounds, but you think that a tape recorder that costs less than 50 pounds is probably not of good quality. You definitely want a tape recorder with auto reverse, and one with a radio

built in would be nice. You have investigated three models of tape recorder and your friend has investigated three models. Get the information from him/her and share your information. You should start the conversation and make the final decision, but you must get his/her opinion, too.
(information about three kinds of tape recorders)

Student B
Your friend is planning to buy a tape recorder, and each of you investigated three types of tape recorders. You think it is best to get a small, light tape recorder. Share your information with your friend, and find out about the three tape recorders that your friend investigated. Let him/her begin the conversation and make the final decision, but don't hesitate to express your opinion.
(information about three kinds of tape recorders)

This kind of task would be evaluated using a system of band scales. The band scales would emphasize the testee's ability to give and receive information, express and elicit opinions, etc. If its intention were communicative, it would probably not emphasize pronunciation, grammatical correctness, etc., except to the extent that these might interfere with communication. The examiner should be an observer and not take part in the activity, since it is difficult to both take part in the activity and evaluate it. Also, the activity should be tape recorded, if possible, so that it could be evaluated or re-checked later.

Role play. In a role play, the testee is given a situation to play out with another person. The testee is given advance information about what his/her role is, what specific functions he/she needs to carry out, etc. A role play task would be similar to the above information gap activity, except that it would not involve an information gap. Usually the examiner or a confederate takes one part of the role play. The following is an example of a role play activity.

Student
You missed class yesterday. Go to the teacher's office and apologize for having missed the class. Ask for the handout from the class. Find

out what the homework was.

Examiner
You are a teacher. A student who missed your class yesterday comes to your office. Accept her/his apology, but emphasize the importance of attending classes. You do not have any extra handouts from the class, so suggest that she/he copy one from a friend. Tell her/him what the homework was.

Again, if the intention of this test were to test communicative language, the testee would be assessed on his/her ability to carry out the functions (apologizing, requesting, asking for information, responding to a suggestion, etc.) required by the role.

Testing Reading and Writing

Some tests combine reading and writing in communicative situations. Testees can be given a task in which they are presented with instructions to write a letter, memo, summary, etc., answering certain questions, based on information that they are given.

Letter writing. In many situations, testees might have to write business letters, letters asking for information, etc. The following is an example of such a task.

Your boss has received a letter from a customer complaining about problems with a coffee maker that he bought six months ago. Your boss has instructed you to check the company policy on returns and repairs and reply to the letter. Read the letter from the customer and the statement of the company policy about returns and repairs below and write a formal business letter to the customer.
(the customer's complaint letter; the company policy)

The letter would be evaluated using a band scale, based on compliance with formal letter writing layout, the content of the letter, inclusion of correct and relevant information, etc.

Summarizing. Testees might be given a long passage—for example, 400 words—and be asked to summarize the main points in less than 100 words. To make this task communicative, the testees should be given

realistic reasons for doing such a task. For example, the longer passage might be an article that their boss would like to have summarized so that he/she can incorporate the main points into a talk. The summary would be evaluated based on the inclusion of the main points of the longer passage.

Testing Listening and Writing/Note Taking

Listening and writing may also be tested in combination. In this case, testees are given a listening text and they are instructed to write down certain information from the text. Again, although this is not interactive, it should somehow simulate a situation where information would be written down from a spoken text. An example of such a test is as follows.

You and two friends would like to see a movie. You call the local multiplex theater. Listen to their recording and fill in the missing information in the chart so that you can discuss it with your friends later.

Theater Number	Movie	Starting Times
1	Air Head	
2		4:00, 6:00, 8:00
3		4:35, 6:45, 8:55
4	Off Track	

Conclusion

Communicative language tests are those which make an effort to test language in a way that reflects the way that language is used in real communication. It is, of course, not always possible to make language tests communicative, but it may often be possible to give them communicative elements. This can have beneficial backwash effects. If students are encouraged to study for more communicative tasks, this can only have a positive effect on their language learning.

Application Exercise

Decide on a situation in which you would give a communicative language test. What is the purpose of the test? What levels of students do you need to test? Do you need to test many students or few? Do you have resources, time, and technology freely available for the test, or are they limited?

Based on the above situation, make a communicative language test that tests at least two skills. What task would you give the testees? How would you try to make the task realistic? How would you evaluate the results? How are the decisions you made about the test tied to the situation you outlined above?

Section IV
The Results

Chapter 15
Scoring Tests

Key Words

machine grading
hand grading
objective tests
subjective tests
adjusting scores

Introduction

Once you give a test, you have to score it. In scoring a test, it is necessary to make the scoring as objective as possible, even if the test is a subjective one. That means that all the graders are grading using the same standard, or, if there is only one grader, that the grader uses the same standard for all papers. For example, if students have to write a short text, and the graders score it without any scoring criteria, different scorers will probably grade it very differently, depending on what criteria they are using. It is important to establish the criteria so that whoever scores the test, the results will be the same, or at least similar. In this chapter, we will cover different types of grading and how grading can be done most effectively.

Machine Grading

Machine grading is the easiest way to grade and will produce the fewest errors in grading. The most common methods are mark cards or sheets on which students blacken the correct answer. What is important for this method is that the card or sheet is easy to read so that students will not make unnecessary errors. Reading a testbook and marking on a separate sheet may create some errors, so it is important to give clear directions and make the relationship between the answer sheet or card and the testbook clear.

It is also important to provide a soft pencil with an eraser for each student and make sure that all students use the right pencil and that they know to completely erase any answers that they want to change.

It is important to estimate the running cost as well as the cost to purchase hardware and software. Certain commercial programs require specific cards or sheets to use for testing, which is usually much more expensive than mass-produced cards or sheets. If you are planning to process many cards or sheets, it might be better to make your own software and cards on the computers you use. Some schools have invested in good grading machines but are not able to use them because they do not have the budget to buy the cards.

Another possible way to do machine grading is to use a computer for testing. This sounds easy, but it can be very complicated. It is necessary to give very clear directions, demonstrate the procedure, and let students practice before the test. Even so, some students will make errors because they do not understand how to use the computer. Thus this method is very good for fast grading without errors, but it is still possible for students to make errors in answering questions.

A major advantage of machine grading is that grading is very fast, easy and accurate. However, from the viewpoint of analyzing test results, an even greater advantage is that the results are already in electronic files, and they can be processed for statistical analyses with different software. Thus, in order to use machine grading to the best advantage, it is important to use the necessary statistical programs to analyze the test results.

Hand Grading

Objective Tests

In many cases, you hand grade objective tests such as multiple choice tests. In that case, it is a good idea to grade only the first few items for all students and grade the next few items for all students, etc., so that you remember all the correct answers while grading. It is important not to stop in the middle. Also if you make an answer section in the test (for example, at the end of the test) and have the students put all their answers at the same place, rather than having answers are scattered all through the exam sheets, it is easier to grade. (However, this makes it easier for students to cheat on the exam and increases the likelihood that students will make mistakes in the placement of their answers.)

Subjective Tests

Subjective tests are more difficult to find a way to evaluate with some degree of objectivity. Before testing, a description of each level of grading should be established, and it should be tried out on a few tests and revised if necessary. If more than one person is doing the grading, they should compare their results on the first few tests and then periodically compare the results throughout the time they are working. If the teacher of a class is working alone, she/he should grade each question for all students without stopping, periodically refering back to the established standards.

In the case of an entrance exam or another kind of test taken by many people, two graders should score the test so that the scores will be more fair. It is best if the same graders can score the same question on all the tests, rather than different graders scoring the same question or the same grader scoring different questions, in order to make the grading as consistent as possible. Graders should be given a concrete grading scale, and they should have a chance to practice with it. Occasionally it is good to check their reliability and adjust their scoring.

Adjusting Scores

For a class test, the best way to find errors in marking is to return the test to the students with the correct answers and the points assigned for each question. Students can check the scoring and calculate the total scores themselves. This will be good feedback and double checks the grading.

In addition, you should explain the reasons why the correct answers are correct. This makes a good review of the material.

For entrance exams or other large scale tests, other people should be appointed to check the calculation of the total scores. If two graders are scoring, it is important to check how different their scores are, and if they are very different, to have the supervisor give the final scores.

Conclusion

Scoring tests is very important, because if it is not accurate, we cannot know what the scores really mean. Also if the scores are used to select students for any purpose, it is not fair to them if the scoring is inaccurate. The scores will be used for instruction, and if the scoring is inaccurate, they are not useful for giving good instruction.

Application Exercise

Look at the tests that your teacher provides or asks you to provide. Can they be machine graded or must they be hand graded? Are they objective or subjective?

Chapter 16

Graphs and Descriptive Statistics

Key Words

graphs
histograms
polygons
pie charts
descriptive statistics
measures of central tendency
measures of dispersion

Introduction

After you have given a test and have the results, it is useful to be able to describe the results. How, in general, did the students do on the test? How spread out were the results? How does an individual student's score compare with the score of other students and with the class as a whole, and how do the scores of one group of students compare with those of another group of students? Graphs and descriptive statistics can help answer these questions. In this chapter, we will describe different kinds of graphs and statistics and how these can help teachers.

Graphs

Graphs can be used to represent how the students did on a test or how students did on different items in a way that is easy to see. Two useful types of graphs are histograms and polygons.

Histograms

To make a histogram representing how the students did on the test, you place the number of points (or ranges of points) across the bottom of the graph and numbers up the side indicating the number of students.

A bar above each number of points or range of points indicates how many students received that grade. Looking at a graph, it is very easy to see how many students received each score and to get an idea of how difficult the test was for the students. If most of the students are on the upper end of the histogram (that is, they scored high on the test), the histogram is said to be negatively skewed, and the test was easy. Conversely, if most of the students are near the lower end of the histogram, it is said to be positively skewed, and the test was difficult.

Similarly, histograms can be used to compare how students did on different items, you can place the item numbers across the bottom of the graph and indicate how many students got the item correct on the left. This can help in evaluating the difficulty of the items. Items intended to test the same thing can also be put together, to give the teacher an idea of what types of items the students found particularly easy or difficult.

Polygons

Polygons are similar to histograms, except a point is used instead of a bar to represent the number of students who received each grade. The points are connected by lines. The advantage of polygons is that the polygons for different groups or classes can be put on transparencies and imposed on each other. This helps the teacher see how two or more groups or classes are different.

Pie Charts

Pie charts are circles divided up into sections. They can be used to show what percentage of the students received grades within certain ranges. This can show graphically how many students are within each range. These are difficult to do by hand, but they can be done using computer programs.

Descriptive Statistics

While graphs are useful to tell at a glance how a class overall did on a test, it is sometimes necessary to have more specific information, so you need to know how to use descriptive statistics and what they mean. It is often useful to know the central tendency and the dispersion of a group

of test takers.

Measures of Central Tendency

Measures of central tendency indicate, in general, where the scores are centered. There are three ways to calculate the central tendency: the mean, the median and the mode. Because each of these measures have strengths and weaknesses, it is helpful to report all of them in order to get a clearer picture of what the results mean.

The mean. The mean is, in general, the most useful measure of central tendency. It is calculated by adding up all the scores and dividing by the number of scores. The mean reflects all of the scores, but one problem with it is that it will be affected by one or a few scores that are very high or very low in comparison to the others.

The mode. The mode is the most common score. It indicates what score the most students received, but it does not give much more information.

The median. The median is the mid-point score, that is, the score with the same number of scores below and above it. (If there are an even number of scores, the mean of the two mid-point scores is taken.) This score may give a clearer picture of the central tendency of the group if there are a few very high or very low scores, because the median is not greatly affected by these.

Measures of Dispersion

In addition to measures of central tendency, there are measures of dispersion that are used to describe the results of tests. These are the range and standard deviation.

Range. The range is a simple indication of how much the scores are spread out. The range is the difference between the highest and lowest score. It is calculated by subtracting the lowest score from the highest score.

Standard deviation. The standard deviation is a more complex way of describing how spread out the scores are. Essentially the standard deviation is the average number of points each person is away from the mean. It is calculated by going through the following process.

1. Calculate the mean (M).
2. List the scores.
3. Subtract the mean from each score. (X-M)
4. Square each difference between the score and the mean. $(X-M)^2$
5. Add up the squares from #4. sum of $(X-M)^2$
6. Divide the sum from #5 by the number of scores, minus one.
7. Find the square root of the result of #6. This is the standard deviation.

$$SD = \sqrt{\frac{\text{sum of } (X-M)^2}{N-1}}$$

The larger the standard deviation, the more spread out the scores are. Two thirds of the scores are within plus or minus one standard deviation of the mean.

Conclusion

Graphs and descriptive statistics are useful for helping the teacher understand what is happening with tests, for comparing different groups taking the same tests and for comparing individual students to the group as a whole.

Application Exercise

I. Twenty students obtained the following scores on Test A. The highest possible score was 10.

8 2 6 3 5 5 6 7 4 9 1 3 7 6 4 5 4 5 3 5

1. Arrange the scores in a histogram.

2. Work out:
a) the mean
b) the mode
c) the median
d) the range
e) the standard deviation

Answer

1 1 *
2 1 *
3 3 ***
4 3 ***
5 5 *****
6 3 ***
7 2 **
8 1 *
9 1 *

a) the mean	4.90
b) the mode	5
c) median	5
d) the range	8
e) the standard deviation	2.00

II. The same twenty students obtained the following scores on Test B. The highest possible score was 10.

2 2 6 3 3 2 3 8 1 9 1 3 8 4 4 2 4 1 3 9

1. Arrange the scores in a histogram.

2. Work out:

a) the mean
b) the mode
c) median
d) the range
e) the standard deviation

Answer

1 3 ***
2 4 ****
3 5 *****
4 3 ***
5 0

6 1 *
7 0
8 2 **
9 2 **

a) the mean	3.90
b) the mode	3
c) median	3
d) the range	8
e) the standard deviation	2.65

III. Compare the results of two tests, and evaluate them.

Chapter 17
Item Analysis

Key Words

facility value
discrimination index

Introduction

When doing objective tests, it is necessary to evaluate the items to see whether there are problems with them. The facility value (FV) and discrimination index (DI) are two measures that are used to help see whether test items are good. In addition, it is useful to look at the incorrect alternatives and see how they worked. Item analysis is customarily used for multiple choice questions, and we will use multiple choice items in our examples. However, item analysis can also be useful for other types of items. In this chapter, we will explain how to calculate these values and how to use them.

Facility Value

The facility value of an item is the percentage of students who answered the questions correctly. If sixty students take a test, and thirty students got a particular item right, the facility value of that item is determined by dividing thirty by sixty. In this case, the facility value is 50%, meaning that half of the students got the item right.

The FV is a quick way to get an idea of how difficult an item is. Obviously an item with a FV of 2% is very difficult for the group of students who took the test, and one with 95% is very easy. It is difficult to make guidelines about what the FV should be, since it depends on a variety of factors. However, using the facility value of the items from pre-test can help you adjust the test to the difficulty that you want. For

example, if you would like the average score to be 70%, and the average on the pre-test is 50%, for the final version of the test, you can choose items with a higher facility value so that the grades on the test will be higher.

Discrimination Index

The discrimination index is another way of analyzing items in a test. It is more difficult to calculate than the facility value, but it also gives more useful information about the strengths and weaknesses of an item.

Assuming that the test as a whole is valid, the students who did well on the test as a whole should do well on any particular item, and the students who did poorly on the test as a whole should do poorly on that item. If students who did well on the test did poorly on a particular item, there is likely to be something wrong with that item. The DI indicates the extent to which a particular item measures the same skill as the test as a whole. In other words, the DI indicates whether the students who did well on the test as a whole did well on that particular item and whether students who did poorly on the test as a whole did poorly on that particular item.

In order to calculate the DI, you divide the students into three roughly even groups, the top group, the middle group and the bottom group, based on their total scores on the test. The top group and the bottom group must have the same number of students. There are different ways to calculate the DI, but one of the easiest is to calculate the percentage of students in the top group who answered correctly and the percentage in the bottom group who answered correctly and subtract the bottom group from the top group. In other words, divide the number of students in the top group who got the correct answer by the total number of students in the top group, and subtract the number of students in the bottom who got the correct answer divided by the total number of students in the bottom group.

$$DI = \frac{\text{\# of correct answers in top group}}{\text{\# of students in top group}} - \frac{\text{\# of correct answers in bottom group}}{\text{\# of students in bottom group}}$$

Take, for example, a group with ten students in the top group and ten students in the bottom group. Imagine that eight of the students in the top group (.8) answered the item correctly and three students in the bottom group (.3) answered the item correctly. .8-.3=.5. The DI for this item is .5, which is fairly good, because it indicates that top students did better on the item than bottom students.

The highest possible DI is 1.0, which indicates that all of the top students answered the question correctly and all of the bottom students answered it incorrectly. This would be an ideal item, but it is very rare. The lowest possible DI is -1.0, indicating that all of the bottom students answered an item correctly and all of the top students answered it incorrectly. Any negative DI indicates that there is something wrong with the item: For some reason, the better students are being tricked into choosing the wrong answer.

Looking at Incorrect Alternatives

When analyzing items, it is also important to look at the incorrect alternatives and at how many students chose each. Looking at how many students from each group chose each alternative can give important information about how well each alternative is working and can guide the teacher in rewriting the alternatives to make a better test.

If many of the top students chose one incorrect alternative, that alternative might actually be correct or close to being correct, or there might be something about it that makes it distracting to the students who understand the point of the question best. The teacher needs to consider how to rewrite the item so that it does not confuse the best students.

If, conversely, no students, even students from the bottom group, chose a particular alternative, the alternative is being wasted. It is not distracting anyone, and should therefore be rewritten to make it more appealing. Ideally, the students in the bottom group should be spread more or less evenly among the incorrect alternatives.

Examples

Look at the following examples. There are ten students in the top group and ten students in the bottom group.

1. "Who are those for?"
   "I'm going to send _____."
   a. my cousin them
   b. to my cousin them
   c. them my cousin
   *d. them to my cousin

	a	b	c	d
Top	1	0	1	8
Middle	2	1	3	4
Bottom	2	3	3	2

Fourteen of the 30 students got the correct answer, so the facility value of this question is 47%. It is a moderately difficult item.

Eight out of ten of the top group and only two out of ten of the bottom group answered the item correctly, so the DI is .6. This indicates that the item is discriminating well between the students who did well on the test and the students who did poorly. In addition, the chosen answers were spread out among the incorrect alternatives, so that at least four out of the thirty students chose any alternative.

2. "Why isn't Deborah coming to the party?"
   "She said she _____ tonight."
   a. has to work
   *b. had to work
   c. have to work
   d. hadn't to work

	a	b	c	d
Top	6	4	1	0
Middle	4	2	4	0
Bottom	2	7	1	0

Thirteen of the thirty students answered this item correctly, giving it a facility value of 43%.

Seven of the bottom students answered correctly, but only four of the top students. The item would have a facility value of -.3, indicating that it is not a good item, because the bottom students did better than the top students. If you look back at the item, you can see that the alternative which was chosen by many of the top students would also be considered correct. It may be that the textbook that the students have used gives the rule that the tense of the verb in the reported speech has to agree with the tense of "said." The better students answered the item incorrectly because they know that that is not necessarily true. The poorer students just went by what the textbook said.

In addition, looking at the way the answers were distributed among the alternatives, it is obvious that D is not a good alternative. None of the students chose it, so they were not fooled by it.

Other Types of Items

A discrimination index can also be calculated for tests when their answers are expressed in words, phrases, or sentences, as long as the answers are either right or wrong and no partial credit is given. However, in that case, there are no distracters to analyze. It is still possible to compare the number of correct answers from the top students and the number of correct answers from the bottom students. It is also informative to study the incorrect answers, particularly those which come from the students in the top group, in order to try to understand why they made mistakes. If their mistakes appear to come from a misunderstanding of the text or from what was expected of them, then the questions should be revised.

Though subjectively-marked items such as essays need to be pretested, item analysis is not appropriate for them.

General Comments

It is difficult to give firm guidelines for what is an acceptable DI. In general, obviously the higher the DI, the better. Anything over .4 is generally considered acceptable.

In using the discrimination index, it is necessary to keep in mind that what it reflects is the test as a whole. If the test as a whole is poorly constructed, then the DI is not really meaningful, since it depends on the validity of the test as a whole. Also, the more items that the test has, the more reliable the results are. If there are few questions, one or two bad items can affect the ranking of the students, making the item analysis suspect. If the test has few items, or if it has not been pretested yet, one solution is to use some outside measure of the students' proficiency rather than the results of the test that is being analyzed. For example, the students could be ranked according to their grades on a standardized test, and this ranking can be used as the basis for putting the students in top and bottom groups.

Conclusion

It is not enough to write items that seem to be good. Items must be pretested to see how students do with them, and to determine the difficulty of the items and how students at different levels do on them. This information can be used to make better items.

Application Exercise

We gave a fifty item test to twenty students. The following are the results of first eight items, and the list is in the order of ranking.

Correct Answer	B	B	C	B	C	D	B	C		total
Question #	1	2	3	4	5	6	7	8	grade	score
1	B	B	C	B	C	D	B	C	8	90%
2	B	C	C	D	C	C	B	C	5	84%
3	B	B	C	D	C	C	D	A	3	72%
4	C	C	C	B	C	C	D	A	3	62%
5	B	C	C	D	B	C	D	A	2	56%
6	B	B	C	D	B	C	D	C	4	54%
7	B	C	C	D	B	C	D	C	3	52%
8	B	C	C	D	B	D	D	A	3	50%
9	B	B	C	B	A	C	D	A	4	48%
10	A	A	C	D	A	A	D	D	1	42%
11	C	C	C	B	A	C	A	A	2	40%
12	D	B	C	C	B	C	A	A	2	36%
13	B	C	C	D	C	B	B	C	5	34%
14	A	B	D	B	A	D	C	D	3	30%
15	B	B	D	B	B	D	A	C	5	28%
16	C	B	B	B	C	A	C	D	3	18%
17	A	B	C	B	D	A	A	A	3	16%
18	B	A	B	B	B	D	D	D	3	14%
19	B	C	C	D	A	C	A	B	2	10%
20	B	B	D	D	A	D	A	B	3	10%

1. Make a chart showing, for each question, the number of students in the top, middle, and bottom group who chose each alternative. Compute the Facility Value and Discrimination Index. (The answers are in the following chart.)

Question #		A	B	C	D	blank	total	FV	DI
	T		6	1			7	86	
1	M	1	3	1	1		6	50	.29
	B	2	4	1			7	57	
	T		3	4			7	43	
2	M	1	2	3			6	33	-.29
	B	1	5	1			7	71	
	T			7			7	100	
3	M			6			6	100	.71
	B		2	2	3		7	29	
	T		2		5		7	29	
4	M		2	1	3		6	33	-.43
	B		5		2		7	71	
	T		3	4			7	57	
5	M	3	2	1			6	17	.43
	B	3	2	1	1		7	14	
	T			6	1		7	14	
6	M	1	1	3	1		6	17	-.43
	B	2		1	4		7	57	
	T		2		5		7	28	
7	M	2	1		3		6	17	.29
	B	4		2	1		7	0	
	T	3		4			7	57	
8	M	4		1	1		6	17	.43
	B	1	2	1	3		7	14	

2. The following are the questions for the first eight items. Using the facility values, discrimination indexes and item analyses, evaluate the questions.

1. Tom read the assignment last week, but I _____.
   A) have read it already
  *B) haven't read it yet
   C) did not read it yet
   D) have read it yet

2. "When are you coming home?"
   "_____."
   A) Last Sunday
  *B) Before dinner
   C) Tomorrow
   D) Yesterday

3. "Do you always leave work this early?"
   "No, _____ I have a date tonight."
   A) if
   B) so
  *C) but
   D) because

4. "She called the doctor three times last week."
   "_____?"
   A) Has she
  *B) Did she
   C) Called she
   D) Didn't she

5. "Is it noon yet?"
   "No, it isn't. It's _____."
   A) a quarter past
   B) eleven o'clock
  *C) a quarter to
   D) ten past

6. "_____ in London is very expensive."
   A) Stay
   B) To staying
   C) To stay
   *D) Staying

7. "_____ a stick of gum?"
   "No, thanks."
   A) Do you have
   *B) Have
   C) You have
   D) Have you got

8. "Must I finish homework tonight?"
   "No, you _____. You can do it tomorrow."
   A) mustn't
   B) can't
   *C) needn't
   D) won't

Chapter 18

Computer Programs for Analyzing Test Results

Key Words

spreadsheets
Iteman
SPSS
SAS
databases
test generating programs
word processor
e-mail
lists

Introduction

Analyzing test results can be very cumbersome, especially if there are a lot of items or a lot of testees involved. There are computer programs available to help teachers analyze the results. Many schools already have statistical packages, and they are very useful for statistical analyses. In this chapter, we will discuss a few of these types of computer software.

Using computer software is not necessarily difficult. If you have colleagues who know how to use a program and are willing to help you, you can learn to use most programs fairly quickly. You will get a printout or a file of the results—but in order to use the results effectively, you need to understand what they mean.

Another problem might be inputing data for analyses. You need to punch in all the scores or other data that you are going to analyze. If you grade with a machine, you can store the results on file, and you can use it for data. If you punch in data, you need to be careful to input data cor-

rectly. It is very easy to make mistakes. It is necessary to proofread the entered data. Nowadays, many programs can read in data from the text made by other programs, and you can punch in the data with the program you are most comfortable with. You should figure out the best way to put data into the software you use for the analyses.

When using statistical packages, be sure that you know what you want to do. It is easy to do any calculation, but it is very easy to lose track of what you are really doing. Make sure that you understand the meaning of the results that you get.

Such basic descriptive statistics as the mean, mode, median, standard deviation, range, negative or positive skew, histogram, etc., are generally fairly straightforward to interpret. However, when you run such statistics as t tests, correlations, regressions, ANOVA, etc., you need to make sure your sample meets the assumptions required by the test and follow correct procedures to check and read the results correctly. Also it is important to be careful to express the results accurately.

Types of Programs Available

Spreadsheets

The easiest program to use is probably a spreadsheet. It can be used to calculate the mean, standard deviation, etc., and to make histograms and pie charts. If you need only descriptive statistics of your test results, probably this is good enough. Microsoft Excel is the spreadsheet that comes with Windows.

Iteman

There are a number of commercially available programs for analyzing test results, and these are useful to have if you need to analyze results and if you understand the information that the program gives you. One item analysis program is called "Iteman." It provides basic descriptive statistics, such as the mean and standard deviation, and information on the reliability of the test. In the case of a multiple choice test, it can provide information about difficulties with incorrect alternatives, for example, when an incorrect alternative has a higher correlation with the

overall results than the correct answer does. (This might indicate that the better students are somehow fooled by a particular alternative, and you need to reconsider it.) Iteman can also compare the responses of upper and lower groups to identify which items are functioning poorly. Iteman is best suited to questions for which there are right or wrong answers, but it can be used for multi-point items with up to nine points. There are standard and extended versions of Iteman, with the standard version able to handle 30,000 examinees and 250 items, and the extended version able to handle 750 items with an unlimited number of examinees.

Iteman is available through Assessment Systems Corporation, 2233 University Ave., Suite 200, St. Paul, MN 55114 USA. They can be contacted by e-mail at mailbox@assess.com, by phone at 612 647-9220, or by fax at 612 647-0412. The cost is approximately $199 for the standard version, $249 for the extended version, and $299 for the Windows version, plus shipping and handling. Information is also available at http://www.assess.com/iteman.html.

SPSS and SAS

SPSS and SAS are available on Windows. They are very powerful, and you can do almost all statistical analyses with them. However, you need to understand statistics first. Without knowing statistics, you cannot use them. If you do use them without understanding statistics, you will probably misinterpret the results that you get.

Databases

Database programs are very useful for keeping test banks. You can store all the items that you have made, and, you can sort them according to whatever category you find useful, including their difficulty level, what skill they test, what grammar point they test, etc. If you keep this database up to date, you can pull out the questions you want in the order you want them to make a test.

Test Generating Programs

There are some commercial programs to make a test with some data they provide or you provide. If you find one which you like, it will be very useful.

Word Processors

Word processors are not as good as databases or test generating programs for storing tests, but if the other types of programs are not available, they are useful for writing, revising, and storing test items.

E-mail

Making a good test is difficult and time consuming. It is a good idea to share work with your colleagues or friends. If you use e-mail, you can save much time and energy exchanging questions, reading and commenting on other people's questions, etc. E-mail makes it possible to easily exchange ideas with people you are working with.

Lists

There are computer discussion lists for the discussion of subjects related to testing and evaluation. LTEST-L (Language testing research and practice) is the largest, and you can subscribe to it by sending the following e-mail message to listserv@psuvm.psu.edu:

subscribe ltesl-l firstname lastname
set ltest-l digest
SIGNOFF

TESL-L is a general list for the discussion of language teaching, but testing and evaluation are sometimes discussed on the list. You can subscribe to it by sending the following e-mail message to listserv@cunyvm.cuny.edu:

subscribe tesl-l your firstname lastname
set tesl-l digest

Conclusion

Technology, particularly computers and their software, are changing every day. Whatever we write here will be somewhat out of date by the time when you read it. Thus, it is important to watch out for what is available and what you can do with it. If you use computer resources wisely, you can save a lot of time and energy, and also you can do things you cannot do without them. In addition to using software to evaluate

BEST COPY AVAILABLE

test results, you can use the Internet to make contact with other teachers to get help and subscribe to computer discussion lists. However, you need to be aware of what you are using and what its limitations are, or you will be wasting rather than saving time and energy.

Application Exercise

Try out one of the programs described in this chapter, or another one that is available to you. What are the advantages and disadvantages of the program? Do you think you would use it as a teacher?

Chapter 19

Evaluating Test Results

Key Words

evaluating results
giving feedback
evaluating instruction
improving tests

Introduction

After tests are administered, they are scored, and the scores are given back to the students. In many cases, that is the end of the process. In some cases, the scores are not even given back, and they are only used for the final class grades. Is this a good use of the testing process? Is the only purpose of tests to get scores or to assign final grades? If so, why is so much time and energy spent on making good tests and administering them? We need to go back to the beginning. Why do we give tests? If we do it to improve instruction and to give feedback to students about their progress, we need to use test results effectively for those purposes. In this chapter, we will discuss how teachers can do this.

How to Evaluate Test Results

As soon as we obtain test results, we need to evaluate them. The most important question is, of course, whether the test has tested what we wanted.

The first thing to do is to calculate the mean and, if possible, the standard deviation. If it is not possible to calculate the standard deviation, one alternative is to make a histogram to show the dispersion of scores. Was the test too difficult or too easy? Are the scores dispersed well? Is the range large enough? These are among the first things you can check.

You should also check to see whether students have understood the questions. Are there any questions which are too difficult or too easy? If there are, you need to consider why they are too easy or too difficult.

As a general rule, an item is too easy or too difficult if the score of question is more than 20 percentage points different from the mean score of the test—but this is only a general guideline. A question that is too easy by this guideline might still be included if, for example, it covers a major concept from the course or if you want to keep some easy questions as an encouragement to lower level students.

If a question is too difficult, you need to try to figure out why. It is possible that for some reason the students did not understand the question. You should reread the question again carefully. Is it somehow tricky or misleading? Is it possible that the testees did not understand what they were expected to do or how they were expected to respond? It is possible to ask colleagues whether the questions seem ambiguous, or even to discuss with students how they interpreted the question. The type of answers students gave may also give an indication of whether they misunderstood the question or what they were expected to do. Do their answers indicate that they were responding to a different question from what you thought you wrote? If it does appear that there was some misunderstanding, you need to consider whether the question should be eliminated or whether there is a way the question can be written in a way that is clearer.

If the mean score is between 65 and 85%, and the scores are well distributed, the overall results are good. If each item is about the same difficulty level and has meaningful wrong answers, they are good. If some questions are too difficult or too easy due to problems in the way the questions were written, you need to revise them.

Giving Feedback to Students

After checking all the questions, you need to decide which questions, if any, to eliminate and then rescore the test, as well as revising for future tests.

Tests should be returned to students with correct answers and some comments. Those comments include what each question or section is

testing, how to study to cover them, as well as what the distribution of scores was for the class, so that students will know how well they did on the test in comparison with other students. You can also indicate above which level scores are particularly good and below which level scores indicate problems. You should encourage students to come to talk about their problems if they got poor grades so that they will learn from their mistakes and not repeat the same mistakes.

It is important to remember that the test is not the final evaluation itself, but part of the process of instruction, and tests should help students learn. This is probably the most important role of tests in instruction.

Evaluating Instruction

Teachers should also use test results to improve instruction. First of all, you should check the overall scores. If the scores are too low, for example less than 50%, the instruction may not have been effective. You should check to see whether students have learned the important points covered in the class. (If more than 80% of the students got an item correct, that item was well covered in instruction.) If students appear to have missed the important points, these points should be reviewed or at least mentioned when the test scores are returned.

In some cases, items on which students do poorly are about vague, abstract ideas. If that appears to be a problem, you need to think about how to teach these points more concretely. Perhaps some of the points have been presented only orally and should be presented in writing.

Improving Tests

You can also use the results of a test to improve that test, using such information as difficulty level, item analyses, problems, etc. That way, you can build up a test bank of tests and test items that have been used with previous students and improved. Writing good test items takes time and energy, and it is a good idea to make the best use of the items you write, improving them and reusing them. Fortunately, word processors make this process easier today, and it is not difficult to keep tests, revise them, and pull them up when you need them.

Any test items that are too easy or too difficult may have to be revised. If an item is too easy, many teachers may be tempted to try to replace it with tricky questions. However, that is not a good idea. Tests should not trick students but should remind them of important concepts covered in the class and see how well they have mastered those concepts. Tricky questions may discriminate, for example, between students who read or listen carefully and those who don't, but that is not the purpose of language proficiency tests.

If students get very low grades on items related to one concept, and the question seems to be too difficult though it is otherwise a good one, you should consider making it easier and keeping both versions, the easier one for lower level students and the more difficult one for advanced students.

In some cases, students make mistakes not because they have not mastered the concept but because they have misunderstood the question or how they are expected to respond. In that case, the item is not really measuring students' language proficiency. If a question appears to be misleading students, you need to rewrite the question more clearly and concretely, so that all the students can understand it.

Conclusion

The results of tests are useful not only for assigning the grades to students, but also for evaluating the tests themselves. If the questions are too easy or too difficult, you should check whether they are good questions or not. If they are not, you need to decide whether to include them in the final score. Even if you decide to include them, you need to revise them for the future tests. Checking the test results carefully gives good feedback to students, and helps teachers to evaluate instruction and give better tests in the future.

Application Exercise

Look at the test provided by the teacher and the results. What can you learn from the results? How could the test be improved? What kind of feedback could be given to the students?

Chapter 20

Reporting the Results of Examinations

Key Words

audiences
teachers
students/parents
administrators
general audience
specialists

Introduction

After you administer examinations, you, of course, grade them and get their results. The next step is to make use of the results. As we have discussed, the results of achievement or progress tests should be used to improve your students' learning and your teaching. However, there are other kinds of examinations, including proficiency tests, entrance examinations, diagnostic tests, placement tests, and aptitude tests. What do you need to report? In the cases of these tests, to whom do you need to report test results? In this chapter, we will discuss what to report and how to report the test results.

General Considerations

You need to make your reports of test results useful to the people who will be reading them, depending on what their purposes are. You need to be aware of the purposes that the report will be put to and also need to present the information in a way that is easy to understand. If, for example, a report is written for teachers, it should be concrete and easy for them to understand. If the teachers use the report to give guidance to individual students, the report should include grades and comments for

individual students. If, on the other hand, the report is intended to give the teacher an overall idea of how the class is doing, it may not need to include individual students' grades and comments, but it should have statistics of central tendency and dispersion to describe the overall results and comments about the performance of the class as a whole. If the tests were given to improve instruction, the reports should include some comments on the areas in which students have weaknesses. In the following section, we will discuss in more detail the types of information useful to different groups of people.

As we have discussed, how you write a test report—which information is included and how you organize the information—depends on the purpose of the report and the intended audience. You need to consider how you can make the test report as useful as possible for its purpose. If the test is a commercial one, you may need to include information on how the test is made, administered, and graded, and information about its reliability, validity, and usefulness. You should also consider what information is useful and necessary—and, nearly as important, what is not. You should carefully consider what to exclude as well as what to include, so that the readers get the information that they need without being distracted by unnecessary information.

Many people who read test reports do not have much knowledge of testing or statistics, so you need to explain how to read and interpret numbers you give in the report. However, you should not explain them too technically, because that makes the explanations difficult to understand and probably tells the readers much more than they really want or need to know. Graphs or charts often help people understand more easily and better than numbers presented in the tables.

On the other hand, if you are reporting to scholars in the area of language testing or reporting the results of tests are a part of your research, you need to report the results in much greater detail. In this case, you need to use technical expressions to make your information clear to specialists. You need to use tables to show detailed information as well as charts and graphs which show the overall results.

In any test reports, you need to clearly explain how the scores are

computed. Are the results a raw score? How were scores weighted for each question? Were any questions eliminated for scoring purposes? If so, why? Were scores based on norms of test takers, or modified using previous test results? Test scores are most important, and they have to be reported clearly so that readers can understand what they mean.

What Do You Report to Different Audiences?

There are five categories of readers for the test results: teachers, students and their parents, administrators of schools or programs (teachers who are not teaching those students), general audiences, and specialists in testing. The people in each of these categories are likely to have different purposes for reading reports on the results of testing, different amounts of background knowledge about language testing, and different expectations about what kind of information they will find in the report.

Teachers

Teachers are concerned with such matters as how well their students are doing compared with other students, what their students' strengths and weaknesses are, how well their students have learned, and how well their students will be able to perform on future tasks, such as entrance examinations or in future classes.

Teachers are interested in specific problems students have, so it is useful to report item analyses and provide the test questions themselves. You can discuss items with which many students had problems and perhaps speculate on why the students might have had a problem. This will help teachers understand what questions were difficult or what problems students had.

You should also give the mean score, standard deviation, and percentile scores of the whole group, each teacher's students as a group, and individual students' scores. This helps teachers understand how well their students did as a group or as individuals.

If teachers are planning to use the test results to counsel individual students, they need to receive the results of each student. The results should be organized in a way that teachers can easily find and under-

stand the results of individual students.

Students and their Parents

Students and their parents are interested in knowing how well the student is doing in comparison with other students and also knowing what the student's problems are. It is good to show where they are ranked, and distribution of whole group, with some explanation to show how much they could move up with how many points.

It is, however, probably more important to show their weakness and, if possible, how they can improve their proficiency. Of course, the latter is also the teachers' responsibility, but in many cases, students do not get much guidance from their teachers, so it is good to include some instructional information, but not too much.

If students are in junior high or below, test results have to be made very easy for them to understand, so they should be given an easier version than their parents are. For high school and university students, test results for students and parents could be the same.

For parents, any information about how they can help students can be useful. For example, if the student is having a serious problem keeping up with the class, he or she may need extra instruction to pass the entrance examinations, etc.

In any case, students and their parents are not likely to be familiar with testing, statistics, or English instruction. Technical terms should be avoided, unless an easy definition is included, and explanations should not be too technical. Less information is necessary about the test itself, how it is administered, etc.

Administrators

Administrators may not know the details of English language teaching. They are interested in selecting or placing students. It is not important to describe in detail individual student's weaknesses in English or problems they have. However, it is important to explain what the scores mean, how students could be classified, and what the reasons are for those classifications. They need concrete standards to decide cutoff points.

It is important to give general objectives of the test, and information

about reliability and validity. Entrance exams in particular are used as a measure for comparing students' achievement and proficiency in English, so it is very important to convince administrators that the test is actually measuring them.

General Audience

Reports on test results are not usually read by general audiences. Only the results of entrance examinations and standardized examinations such as TOEIC and TOEFL might be read by general audiences, since they are released to the public.

The test results should clearly show the purpose and intended test takers and how results are analyzed for that purpose. It is important to show how reliable and valid the test is, and for what purpose it is a good test. It is also important to show how fairly the test was administered and graded. For that purpose, the grading procedures need to be explained clearly. In addition, if the test is given periodically, it is necessary to report how the results of each test are adjusted.

It is important to make the meanings of the scores clear as well as how they can be used for different purposes, so that the results will not be used in ways they were not intended to be used.

Specialists in Testing

Ordinary teachers seldom have chances to report test results to specialists. However, if you are engaged in research, you will need to write a report or paper for specialists in testing. In order to write good research reports or research papers, you need to read ones that others have written so you can get an idea of the conventions that they use. However, the most important point is that you need to report your methods and results specifically and clearly, so that other researchers understand what you did and can follow the same procedures.

You should state the purpose of the test and its specifications clearly and explain how the test items were made or selected, how they were pre-tested, how they were finalized, how they were administered to the students, and how they were scored. As for the results, you need to provide not only descriptive statistics but also statistics indicating the

reliability of the test itself and compare it to other standardized tests, if possible.

You should report everything that readers need to know, not just the information that supports your hypothesis, if you have one. You need to report how you analyzed the data, the results of your analyses, and your interpretations. You should honestly discuss the weaknesses of your study.

Conclusion

It is important to make good tests and to administer and grade them properly. However, reporting test results well enough for people who use them is also very important, because they can use them to learn or teach English better, improve classes, and select good students. If the reports of test results are not good, people may not use them, or they may misunderstand them and use them in a way they were not intended to be used. In those cases, we waste time, energy, and money.

Discussion Questions

Look at the results of a test that your teacher gives you. Discuss how they might be presented differently to different audiences.

Appendix A

Types of Language Testing Resources on the Internet

Keywords

web pages
lists of links
organizations
computer discussion lists/mailing lists
publications and articles
standardized tests

Introduction

The World Wide Web has a large number of sites related to language testing, and you can find many useful resources. In this appendix, we will discuss some of the types of sites that are available, along with information about specific sites. (You can also find information at our web site "Language Testing" [http://ilc2.doshisha.ac.jp/users/kkitao/online/www/test.htm] This is one section of "On-Line Resources and Journals: ELT, Linguistics, and Communication" [http://ilc2.doshisha.ac.jp/users/kkitao/online/].)

Language Teaching Resources

The Internet has a great deal of information about testing and a great many test-related resources. Among the materials are web pages and information about organizations and universities, mailing lists related directly or indirectly to testing, web pages of publications, articles and papers, information about English language tests used internationally, and information about non-English language tests.

The resources include web pages with lists of links related to language testing. Among the best is "The Resources in Language Testing WWW Page" (http://www.surrey.ac.uk/ELI/ltr.html). It has an enormous amount of information about testing, including many links to organizations, publications and tests. It also have reviews of pages that it has links to, so you can find out which web pages would be likely to be most useful to you.

Yahoo has a page of annotated links to pages related to testing at http://dir.yahoo.com/Education/Standards_and_Testing/. Some of the pages are related to testing in general and some to English language tests. The CALL Lab at the University of Ohio maintains a links page to information about language tests and practice tests called "Language Tests for English Language Learners" (http://www.tcom.ohiou.edu/OU_Language/english/tests.html).

Many organizations that administer English language tests, professional organizations related to language testing, and universities which have special projects or activities related to language testing have web pages. You may find some of them useful for obtaining information on various language tests, events, and trends in the profession. Among the major organizational web pages are The Educational Testing Service (http://etsis1.ets.org/), The International Language Testing Association (http://www.surrey.ac.uk/ELI/ilta/ilta.html), and The University of Cambridge Local Examinations Syndicate (UCLES) (http://www.ucles.org.uk).

Computer discussion lists or mailing lists are lists where subscribers discuss a given topic, ask questions, etc. Subscribers do this by sending an e-mail message to a central computer, which distributes the message to all subscribers. If a member of the list wants to respond, he/she replies in a message to the computer, and again it is distributed to all subscribers. LTEST-L is the mailing list most directly related to language testing. You can subscribe to it by sending the message with a blank subject line "subscribe ltesl-l firstname lastname" (without the quotation marks) to listserv @psuvm.psu.edu. (You can find more information about LTEST-L at http://ilc2.doshisha.ac.jp/users/kkitao/online/list/lis-tef2.htm#ltest-l). In addition, other mailing lists related to English language teaching occasionally have contributions on the subject of testing. You can find information about these lists at "Useful Mailing Lists for TEFL/TESL" (http://ilc2.doshisha.ac.jp/users/kkitao/online/list/lis-tefl.htm). The largest mailing list in ESL/EFL is TESL-L, which has a huge archives, including articles about testing. You can find information about subscribing to TESL-L and accessing articles from its archives at http://ilc2.doshisha.ac.jp/users/kkitao/online/list/lis-tesl.htm.

A number of testing-related publications have web pages. Some include articles, and others only include tables of contents and subscription information. Among the major journals are *Language Testing Update* (http://www.ling.lancs.ac.uk/pubs/ltu/ltumain.htm) and *Language Testing* (http://

www.surrey.ac.uk/ELI/langtest/ltsearch.html). ERIC is an organization that has a huge database of articles related to various aspects of education. ERIC Clearinghouse for Evaluation and Assessment (http://www.ericae.net/) has a searchable database of language testing information, along with some articles about testing. (You can search the entire ERIC Database, which includes papers on areas other than testing, at http://ericir.syr.edu/Eric/index.html.) The University of Surry has a list of links to testing-related articles on line (http://www.surrey.ac.uk/ELI/artlt.html).

There are web sites related to specific tests, including Cambridge ESL/EFL Examinations (http://www.ucles.org.uk/), IELTS (http://www.ielts. org/ and http://www.britcoun.org/eis/ielts.htm), TOEFL (http://www.toefl. org/ and http://www.testwise.com/review.html), and TOEIC (http://www.toeic.com/ and http://www.aitech.ac.jp/~iteslj/Articles/Gilfert-TOEIC. html).

In addition, there are pages that give information about or links to a variety of tests, such as "English Language Examinations" (http://www.ilcgroup.com/ILC_ExamsBC.html), "Dave's Quiz Directory" (http://www.eslcafe.com/search/Quizzes/), "Grammar and Vocabulary Tests and Quizzes" (http://dwjc.dwc.doshisha.ac.jp/kkitao/quiz.htm), "Interactive English Language Exercises" (http://www.ilcgroup.com/interactive/), and "Quizzes" (http://ilc2.doshisha.ac.jp/users/kkitao/class/material/quiz/).

If you need to review previous publications for your studies or research, "Bibliography on Language Testing" (http://ilc2/users/kkitao/library/ biblio/test-bib.htm) may be useful to you. It is fairly complete up to summer, 1995, and it is useful if you are looking for publications.

Conclusion

If you want to find information about testing, the Internet is an excellent resource. It has information about professional organizations, information about specific standardized tests, discussion lists, and so on. Appendix B of this book has more links to Internet resources related to language testing.

Application Exercise

Look at the web pages discussed in this appendix. Which do you think would be most useful to you? Why?

Appendix B

A List of Language Testing Resources on the Internet

The following is a list of Internet resources related to language testing. (This is an adaptation of a web page that can be found at http://ilc2.doshisha.ac.jp/users/kkitao/online/www/test.htm.) The resources are divided into six sections—"Resources," "Organizations and Universities," "Mailing Lists," "Publications," "English Language Tests," and "Tests of Non-English Languages."

In the cases of web pages, the URL (address) and title of the web page is given. In some cases, there is an additional description. If the description is unmarked, it is from the web page. If the description is marked with two stars (**), it was added by the authors.

1. Resources

http://www.ed.gov/databases/ERIC_Digests/ed385609.html
Assessment & Evaluation on the Internet
**Links to web pages of organizations, etc., with annotations.

http://www.mrtc.org/~twright/quizzes/quizcenter/
ED TECH TOOLS Free On-Line Services for Educators
The Ed Tech Tools are designed to assist educators in bringing technology to the classroom. They aim to use the World Wide Web in a greater capacity; to extend its capabilities further than a display of text and graphics.

http://dir.yahoo.com/Education/Standards_and_Testing/
Education Standards and Testing
**Links to web sites related to testing.

http://www.cal.org/CAL/HTML/FLTests.htm
Foreign Language Test Development (Speaking Tests)

http://www.surrey.ac.uk/ELI/ltrconf.html
Forthcoming Conferences

http://grove.ufl.edu/~klilj/#testeval
Kristin's ESL Page ESL Testing and Evaluation sites

http://www.tcom.ohiou.edu/OU_Language/english/tests.html
Language Tests for English Language Learners

http://www.surrey.ac.uk/ELI/cybertxt.html
The Language Tester's Guide to Cyberspace
**Links to web sites related to language testing, arranged by type.

http://www.actfl.org/
Oral Proficiency Testing ACTFL

http://www.surrey.ac.uk/ELI/ltr.html
The Resources in Language Testing WWW Page
The Resources in Language Testing Page was first published on 1 May 1995, and this frames version of the site was launched on 20 July 1996. Its purpose is to act as a reference guide to language testing related resources on the Internet. At this site you will find all the links which have been collected to date, and each one is reviewed in the Language Tester's Guide to Cyberspace. Instead of randomly visiting sites, you can select those which are most likely to provide the information you are looking for.
**This is very comprehensive site, and you should definitely visit it.

http://gwis.circ.gwu.edu/~eaceast/products/proftest.html
Tst Library Inventory Evaluation Assistance Center East Language Proficiency Tests

http://www.wfi.fr/volterre/test.html
VOLTERRE-FR English and French Testing Services

http://138.236.40.190/langtest/index.html
W3 Foreign Languages Testing Database

TESL-L Archives

If you are subscribed to the TESL-L mailing list, you can retrieve a large number of articles from their archives, including articles on testing. See the section on mailing lists for a link to information about joining TESL-L.

Once you subscribe to TESL-L, you can receive a list of testing-related articles by sending the message "index testing" (with a blank subject line) to listserv@cunyvm.cuny.edu. (See http://ilc2.doshisha.ac.jp/users/kkitao/online/www/test.htm#tesl-l for more detailed information about how to retrieve articles from the TESL-L archives.)

2. Organizations and Universities

http://www.alte.org/index2.htm
ALTE The Association of Language Testers in Europe

http://www.cambridge-efl.org.uk/exam/index.htm
Cambridge EFL Examinations

http://www.lsa.umich.edu/eli/testing.html
English Language Institute (ELI) Testing-Certification

http://etsis1.ets.org/
ETS
**Links related to Educational Testing Service, its tests, and links to other test-related resources

http://www.sfsu.edu/~testing/testagnt.html
ETS (Testing Agencies)
**Links to information about GRE, GMAT, PRAXIS, SAT, and TOEFL

gopher://gopher.ets.org/11/
ETS gopher
gopher://gopher.ets.org/
ETS gopher site

http://www.man.ac.uk/IATEFL/iatefl.html#11
IATEFL: Testing Special Interest Group
**Description of the special interest group

http://www.surrey.ac.uk/ELI/ilta/ilta.html
ILTA International Language Testing Association

http://www.ling.lancs.ac.uk/groups/ltrg/index.htm
Lancaster University Language Testing Research Group

http://www.warwick.ac.uk/fac/soc/CELTE/testunit.html
The Language Testing and Evaluation Unit
**This unit of Warwick University provides supervision for research students and professional test consulting services.

http://www.arts.unimelb.edu.au/Dept/LTRC/LTRC.html
LANGUAGE TESTING RESEARCH CENTRE The University of Melbourne
**Information about the center, its staff, its publications, etc., plus links to other testing web sites

http://www.ucles.org.uk/
University of Cambridge Local Examinations Syndicate (UCLES)

3. Mailing Lists

LTEST-L (Language testing research and practice): listserv@psuvm.psu. edu
subscribe ltesl-l firstname lastname
set ltest-l digest
SIGNOFF

Ltest-l is an open forum for discussion on issues pertaining to language testing theory and research. The members are mainly university professors and graduate students but there are also researchers in such institutions as the Educational Testing Service. Membership is open to anyone from any country. We have members from Australia, the UK and other European countries. (I'm not sure about Asia.)
Ilze Duarte, list co-manager ibd2@psuvm.psu.edu

http://ilc2.doshisha.ac.jp/users/kkitao/online/list/lis-tesl.htm
TESL-L
**The largest mailing list for TESL/TEFL; sometimes has discussions related to testing.

http://ilc2.doshisha.ac.jp/users/kkitao/online/list/lis-tefl.htm
Useful Mailing Lists for TEFL/TESL

4. Publications

Journals

ERIC Clearinghouse on ASSESSMENT AND EVALUATION (TM)
Catholic University of America
210 O'Boyle Hall
Washington, DC 20064-4035
Telephone: 202-319-5120 FAX: 202-342-5033
Internet: eric_ae@cua.edu

All aspects of tests and other measurement devices. The design and methodology of research, measurement, and evaluation. The evaluation of programs and projects. The application of tests, measurement, and evaluation devices/instrumentation in education projects and programs.

http://www.surrey.ac.uk/ELI/langtest/ltfind.html
Language Testing Database

You can now search for articles in Language Testing by author (e.g., "Alderson"), or keyword(s) in a title (e.g. "tasks," "self-assessment," or "item response theory").

http://www.surrey.ac.uk/ELI/langtest/ltsearch.html
Language Testing Search Engine
**A searchable database of tables of contents from back numbers of Language Testing.

http://www.ling.lancs.ac.uk/pubs/ltu/ltuindex.htm
LANGUAGE TESTING UPDATE Index: Issues 1-15

http://www.ling.lancs.ac.uk/pubs/ltu/ltumain.htm
Language Testing UpDate homepage

Dissertations

http://www.surrey.ac.uk/ELI/testd.html
Dissertations on testing
**Abstracts of MA dissertations related to testing written at the University of Surrey

Papers and Articles

http://www.surrey.ac.uk/ELI/artlt.html
Articles and Reviews related to Language Testing on the Internet
**links to many articles and reviews

http://call-centre.cogsci.ed.ac.uk/CALLResearch/AAC/AACResearchPapers/AssessofLanguageComm
New Educational Approaches to the Assessment of Language and Communication

http://darkwing.uoregon.edu/~ode/archive/v96/3/050295/050295.test.html
"Test may help non-native English speakers" by Sarah Gray
Oregon Daily Emerald, Tuesday, May 2, 1995 Volume 96, Number 144

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/
Articles on Testing by S. Kathleen Kitao and Kenji Kitao (Most of these are in the TESL-L archive.)

http://ietn.snunit.k12.il/newslet5.htm#whytest

Why Do We Test? English Teachers' Electronic Newsletter, No. 5 (March, 1996) [Israel]

An article that discusses why we test students, by S. Kathleen Kitao

http://www.aitech.ac.jp/~iteslj/Articles/Kitao-Testing.html

Testing Communicative Competence The Internet TESL Journal, Vol. 2, No. 5 (May, 1996) [Japan]

S. Kathleen Kitao and Kenji Kitao (1996)

An article on how to test communicative competence.

(ERIC Document Reproduction Service No. ED 398 260)

http://www.aitech.ac.jp/~iteslj/Articles/Kitao-TestingGrammar.html

Testing Grammar The Internet TESL Journal, Vol. 2, No. 6 (June, 1996) [Japan]

S. Kathleen Kitao and Kenji Kitao

How to test discrete grammatical items

(ERIC Document Reproduction Service No. ED 398 256)

http://www.aitech.ac.jp/~iteslj/Articles/Kitao-TestingListening.html

Testing Listening The Internet TESL Journal, Vol. 2, No. 7 (July 1996) [Japan]

S. Kathleen Kitao and Kenji Kitao

How to test listening skills

(ERIC Document Reproduction Service No. ED 398 257)

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/reading.htm

Testing Reading {TESTING READING}

An article that discusses reading skills (16 are listed) and how to test them.

(ERIC Document Reproduction Service No. ED 398 258)

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/speaking.htm

Testing speaking {TESTING SPEAKING}

An article that discusses ways to test speaking skills.

(ERIC Document Reproduction Service No. ED 398 261)

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/vocab.htm

Testing Vocabulary {TESTING VOCAB}

An article on testing vocabulary in ESL/EFL classes.

(ERIC Document Reproduction Service No. ED 398 254)

http://agoralang.com/agora/agoranews/agoranews2.7
Testing Writing Agora Newsletter, Vol. 2, No. 7 (July, 1996)
S. Kathleen Kitao and Kenji Kitao

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/writing.htm
Testing Writing (1996) [TESTING WRITING]
How to test the writing of ESL/EFL students
(ERIC Document Reproduction Service No. ED 398 259)

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design.htm#validity
Validity and Reliability (1996) [TEST DESIGN]
An article on validity and reliability

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design.htm#common
Common Problems with Test Items (1996) [TEST DESIGN]
An article on common problems with test items

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design.htm #item
Item Analysis (1996) [TEST DESIGN]
An article on item analysis

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design2.htm#good
Writing a Good Test (1996) [TEST DESIGN2]
An article on writing a good test

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design2.htm#specification
Test Specifications (1996) [TEST DESIGN2]
An article on test specifications

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design2.htm#standard
What You Need to Learn about Standardized Tests (1996) [TEST DESIGN2]
An article on standardized tests

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/eval.htm # computer
Computer Programs for Analyzing Test Results (1996) [TEST EVAL]
An article on computer programs for analyzing test results

BEST COPY AVAILABLE

http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/eval.htm# report
Reporting the Results of Examinations (1996) ELT News & Views 3(3) Feature Supplement: Exams and Testing.
[TEST DESIGN2]
An article on reporting examination results

Bibliography

http://ilc2.doshisha.ac.jp/users/kkitao/library/biblio/test-bib.htm
Bibliography on Language Testing

http://ericir..syr.edu/Eric/
ERIC Searching ERIC database since 1966.
**You can search for ERIC documents by title, author, keyword, publication year, or ERIC number. There are 957 ERIC documents that have both "language" and "test" in the title. And there are 19,486 documents that have both "language" and "test" as keywords, as of April 24, 1999.

5. English Language Tests (Cambridge, IELTS, TOEFL, TOEIC, etc.)

Various English Language Tests

http://www..ilcgroup.com/ILC_ExamsBC.html
English Lamguage Examinations
Cambridge,, Oxford, ARELS and other British examining boards.

gopher://gopher.ets.org/11/
Educationall Testing Service Gopher (Get info on the TOEFL, etc.)

http://www:.sas.upenn.edu/elp/programs/testing.htm
ELP Testing Services
**List of various English language proficiency tests.

http://www:.central-london-english.co.uk/examinations.html
Examinations
**Informatiion about various examinations

http://www:.tcom.ohiou.edu/OU_Language/english/tests.html
Language Tests for English Language Learners Ohio University

http://www:.swan.ac.uk/cals/exam.htm
Public Engllish Language Examinations
**Informatiion about various examinations.

http://www.cambridge-efl.org.uk/exam/
Cambridge EFL Examinations

http://math.unr.edu/linguistics/toefl.links.html
TOEFL and other testing information Linguistic Funland

IELTS

http://www.swan.ac.uk/cals/ielts.htm
Centre for Applied Language Studies--IELTS Examinations

http://www.britcoun.org/eis/ielts.htm
IELTS—English language test
**Information about IELTS, when it is offered, how to apply to take it, etc.

http://www.ielts.org/index.html
IELTS Home Page

http://www.une.edu.au/ltc/ielts.htm
IELTS Testing

http://www.ling.lancs.ac.uk/groups/ltrg/washback.htm
Language Testing Research Group—IELTS Washback Project

TOEFL

http://www.kyoto-su.ac.jp/information/tesl-ej/ej06/mr2.html
A review of the TOEFL Mentor CD-ROM by Thomas Robb

http://www.testwise.com/review.html
Review of the TOEFL Test
**Includes examples of different types of questions.

http://www.jyu.fi/tdk/hallinto/ka/Toefl/Toefl0.html
Test Of English as a Foreign Language
This page is intended to give you an overview what to expect from the TOEFL test.

http://ieie.nsc.ru/~eac/toefl.html
Test of English as a Foreign Language
http://www.eas.asu.edu/~leaders/Tests/toefl.html
Test of English as a Foreign Language (TOEFL)

http://www.schoolnet.ca/snc/esl-snc/slinks.html#TOEFL
TOEFL

http://www.toefl.org/
TOEFL homepage

http://web1.toefl.org/
TOEFL.org

http://www1.kaplan.com/view/article/0,1898,1107,00.html
The TOEFL

gopher://gopher.ets.org:70/11/ets.tests/toefl
TOEFL Gopher Menu

http://www.linguistic-funland.com/toefl.links.html
TOEFL Information on the Internet Linguistic Funland
**list of links

http://www.toefl.org/edtstprg.html
TOEFL Institutional Testing Program

http://www.encomium.com/
TOEFL Mentor
**A CD-ROM test preparation system

http://www.toefl.org/tstprpmt.html
TOEFL practice questions
**Practice questions and test preparation products.

http://studint.net/toefl/
TOEFL Prep

http://www.stuff.co.uk/toefl.htm
TOEFL Practice

http://etsis1.ets.org/shopping/ets/LEVEL1/LEVEL2/TOEFL/leaf/product1. html
TOEFL related material ETS Store

gopher://gopher.ets.org:70/11/ets.tests
TOEFL and TOEIC gopher menu

http://www.studyusa.com/articles/admtest.htm
Where to register in Your Region
Admission Tests for the TOEFL, TSE, SAT, GRE, GMAT, or MAT

TOEIC

http://toeic-usa.com/
International Communications Inc. (U.S.)
**Information about TOEIC.

http://www1.kaplan.com/view/article/0,1898,1251,00.html
"Introduction to the TOEIC"

http://www.aitech.ac.jp/~iteslj/Articles/Gilfert-TOEIC.html
"A Review of TOEIC" by Susan Gilfert
**This paper provides a lot of information about the TOEIC.

gopher://gopher.ets.org:70/11/ets.tests
TOEFL and TOEIC gopher menu

http://www.toeic.com/
TOEIC official homepage

http://www.bekkoame.or.jp/~nerm/toeic.htm
TOEIC

http://ilc2.doshisha.ac.jp/users/kkitao/japanese/library/kyozai/TOEIC/
TOEIC (in Japanese)

http://www.faceweb.okanagan.bc.ca/toeic/
TOEIC Preparation
**Sample TOEIC questions.

Other English Language Tests

http://www.study.com/activities.html
Activities
**Quizzes.

http://www.lsa.umich.edu/eli/aee.htm
Academic English Evaluation
**Describes the test used at the University of Michigan.

http://www.cal.org/CAL/HTML/bestest.htm
Basic English Skills Test (BEST)

http://www.eslcafe.com/search/Quizzes/
Dave's Quiz Directory
**Extensive list of links to quizzes, with annotations.

http://www.acer.edu.au/products/ed_resources/990de.htm
Diagnostic English Language Test (DELTA)
**Description of this test.

http://www.ilcgroup.com/ILC_ExamsBC.html
English Language Examinations Cambridge, Oxford, ARELS and other British examining boards.
This file lists the major English language tests offered by British examination boards which are open to non-native speakers of English.

http://www.pacificnet.net/~sperling/quiz/
ESL Quiz Center
**Links to quizzes in such categories as geography, grammar, history, and science.

http://www.lsa.umich.edu/eli/ecce.htm
Examination for the Certificate of Competency in English (ECCE)

http://www.lsa.umich.edu/eli/ecpe.htm
Examination for the Certificate of Proficiency in English (ECPE)
**Explanations of these tests.

http://dwjc.dwc.doshisha.ac.jp/kkitao/quiz.htm
Grammar and Vocabulary Tests and Quizzes

http://www.ilcgroup.com/interactive/
Interactive English Language Exercises

http://www.aitech.ac.jp/~iteslj/quizzes/js/
Interactive Javascript Quizzes for ESL Students
**A wide variety of different types of quizzes, with annotations.

http://www.lsa.umich.edu/eli/igsi.htm
The International Graduate Student Instructor (IGSI)
**Description of a test for graduate student instructors.

http://www.lsa.umich.edu/eli/melab.htm
The Michigan English Language Assessment Battery (MELAB)

http://ietn.snunit.k12.il/onlineex.htm
Online Exercises and Tests for the EFL Learner

http://quizsite.com/
The quiz site
**You can search for types of quizzes by keyword.

http://ilc2.doshisha.ac.jp/users/kkitao/class/material/quiz/
Quizzes

http://www.geocities.com/RainForest/Vines/8255/index.html
The Quiz Zone

http://grove.ufl.edu/~klilj/samptest.htm
Sample Tests and Methods for Evaluation
**Links to a large number of different sample tests, including tests for vocabulary, grammar, pronunciation, and listening.

http://www.aitech.ac.jp/~iteslj/quizzes/index.html
Self-Study Quizzes for ESL Students
**Content-area and grammar quizzes.

http://carla.acad.umn.edu/SOPI/SOPI-practice.html
Simulated Oral Proficiency Interview (SOPI)

http://yamabuki.yamabuki-hs.shinjuku.tokyo.jp/jp/org/eng/step.html
The STEP test (in Japanese)

http://www.stuff.co.uk/media/cat/cat.html
Test your level of English

http://www.livjm.ac.uk/language/engtest.htm
Test Yourself: E N G L I S H

http://lib.upm.edu.my/iistes.html
Testing Skills

http://www1.kaplan.com/view/article/0,1898,1136,00.html
The TSE
**Information about the Test of Spoken English

http://www1.kaplan.com/view/article/0,1898,1132,00.html
The TWE
**Information about the Test of Written English

Tests of Non-English Languages

http://www.collegeboard.org/ap/french/
Advanced Placement French Language Examination

http://www.collegeboard.org/ap/german/
Advanced Placement German Language Examination

http://www.collegeboard.org/ap/students/index.html
Advanced Placement Program

http://www.collegeboard.org/ap/spanish/
Advanced Placement Spanish Language Examination

http://ourworld.compuserve.com/homepages/Joschu/
TransWord Foreign Language Tests / Exercises

日本出版権英潮社所有

本書を無断で複写（コピー）することは、著作権法上認められている場合を除き、禁じられています。小社は、著者から複写（コピー）に係る権利の管理につき委託を受けていますので、複写をされる場合は、必ず小社宛ご連絡ください。

英語テスト法概論［英文］
[Essentials of English Language Testing]

1999年10月 1日　初版発行

定価　本体3,000円（税別）

著　者　北尾S.キャスリーン
　　北　尾　謙　治
発行者　加　藤　昌　雄

発行所　東京都千代田区神田神保町2-28
　　日下ビル（〒101-0051）
　　電話 東京（3263）1641・6171（代）

株式会社 英潮社

ISBN4-268-00327-4 C3082 ¥3000E

BEST COPY AVAILABLE

英潮社

BEST COPY AVAILABLE

162

定価 本体3,000円(税別)

ISBN4-268-00327-4 C3082 ¥3000E

Eichosha Co., Ltd., Kusaka Building, 2-28 Kanda Jimbocho, Chiyoda-ku, Tokyo 101-0051